SeaDataCloud

A Web-Based Virtual Research Environment for Marine Data

EGU-Online 2020
Session ESSI2.11 – Earth/Environmental Science Applications on HPC and Cloud Infrastructures

Presented by Merret Buurman, German Climate Computing Centre (DKRZ), Germany
buurman@dkrz.de

Sebastian Mieruch (Alfred-Wegener-Institut (AWI), Germany)
Alexander Barth (GHER, Université de Liège, Belgium)
Charles Troupin (GHER, Université de Liège, Belgium)
Peter Thijsse (MARIS, The Netherlands)
Themis (Greek Research and Technology Network (GRNET), Greece)
Naranyan Krishnan (Science and Technology Facilities Council (STFC), United Kingdom)

What is this about...

We‘re building an **online environment for processing marine data**.

These are mostly existing softwares developed and optimized for Desktop usage.

While porting these software to the cloud-based environment, we had to solve some problems.

In the coming slides, we will present four examples.

The SeaDataCloud project (2016-2020) has received funding from the European Union's Horizon 2020 Research Infrastructure programme, under the grant agreement No 730960.

Deployment across Europe

1 The compute resources are located at computing centres **across Europe** (Greece, Germany, UK, …). But users need to be able to **access their input data** in all of these.

Data access over the Network (WebDAV, NFS, …) tends to be too slow, especially as most softwares are interactive, so users are waiting for data access.

The SeaDataCloud project (2016-2020) has received funding from the European Union's Horizon 2020 Research Infrastructure programme, under the grant agreement No 730960.

Deployment across Europe

1 Our approach:

We compared data access by various services and adapted our deployment:

I. Heavy continuous IO: Deploy **close to data**, direct disk access or via NFS.

II. Read once, write once: Deploy **further away** and synchronize the data at strategic moments.

III. Services whose results are unlikely to be input to other services (due to formats, etc.): Deploy anywhere

(Locking: Difficult to have all services implement this, especially across file systems.

The SeaDataCloud project (2016-2020) has received funding from the European Union's Horizon 2020 Research Infrastructure programme, under the grant agreement No 730960.

File permissions and data sharing

2 The applications run out-of-the-box as different Linux users (**uids**), which leads to problems for **shared data access**. E.g. NextCloud runs as uid 33 out-of-the-box, Jupyter Notebooks run as uid 1000, …

The SeaDataCloud project (2016-2020) has received funding from the European Union's Horizon 2020 Research Infrastructure programme, under the grant agreement No 730960.

File permissions and data sharing

2 The applications run out-of-the-box as different Linux users (**uids**), which leads to problems for **shared data access**. E.g. NextCloud runs as uid 33 out-of-the-box, Jupyter Notebooks run as uid 1000, …

Our approach:

- If they run on separate servers: Change uid during synchronisation.
- Otherwise: Force all to run as uid 33.

The SeaDataCloud project (2016-2020) has received funding from the European Union's Horizon 2020 Research Infrastructure programme, under the grant agreement No 730960.

Authentication & Authorization

3 Desktop applications have no **authentication** and authorization. It has to be **added to their web-GUIs.**

The SeaDataCloud project (2016-2020) has received funding from the European Union's Horizon 2020 Research Infrastructure programme, under the grant agreement No 730960.

Authentication & Authorization

3 Desktop applications have no **authentication** and authorization. It has to be **added to their web-GUIs.**

Two possible ways:

- **Every application** authenticates user.
- Or **centralized**: A central proxy authenticates & authorizes every single request.

The SeaDataCloud project (2016-2020) has received funding from the European Union's Horizon 2020 Research Infrastructure programme, under the grant agreement No 730960.

Authentication & Authorization

3 Desktop applications have no **authentication** and authorization. It has to be **added to their web-GUIs.**

Our approach:

- **Every application** authenticates user.

This is more loosely-coupled. This way, the institutions can also deploy their processing services stand-alone.

The SeaDataCloud project (2016-2020) has received funding from the European Union's Horizon 2020 Research Infrastructure programme, under the grant agreement No 730960.

Container spawning

4 Some applications cannot handle several users and need a **new instance to be started for every user.**

The SeaDataCloud project (2016-2020) has received funding from the European Union's Horizon 2020 Research Infrastructure programme, under the grant agreement No 730960.

Container spawning

4 Some applications cannot handle several users and need a **new instance to be started for every user.**

Our approach:

We use **JupyterHub** (with dockerspawner) to spawn these services. JupyterHub can not only be used for Jupyter Notebooks, but for any containerized service that serves HTTP content at the expected URL.

The SeaDataCloud project (2016-2020) has received funding from the European Union's Horizon 2020 Research Infrastructure programme, under the grant agreement No 730960.

Container spawning

4 **Benefits of JupyterHub**:
It takes care of user authentication and authorization, proxying, user management, SSL termination.

A full orchestration tool such as kubernetes would likelily scale better, but using JupyterHub is a **good option for small deployments**, or quick tests by institutions that want to avoid the effort of getting into k8s.

The SeaDataCloud project (2016-2020) has received funding from the European Union's Horizon 2020 Research Infrastructure programme, under the grant agreement No 730960.

… Thanks for reading!

For questions, don‘t hesitate to ask in the chat, or contact me at buurman@dkrz.de. I am happy to discuss these topics anytime!

The SeaDataCloud project (2016-2020) has received funding from the European Union's Horizon 2020 Research Infrastructure programme, under the grant agreement No 730960.

Funding & License

The SeaDataCloud project (2016-2020) has received funding from the European Union's Horizon 2020 Research Infrastructure programme, under the grant agreement No 730960.