Geophysical Research Abstracts
Vol. 21, EGU2019-16149, 2019
EGU General Assembly 2019

Virtual Research Environment for FAIR environmental data lifecycle management in the Cloud

Fernando Aguilar (1), Daniel García (2), María Castrillo (2), and Jesús Marco (1)
(1) IFCA-CSIC, Santander, Spain (aguilarf@ifca.unican.es), (2) IFCA-UC, Santander, Spain (garciad@ifca.unican.es)

The eXtreme-DataCloud project (XDC), under the umbrella of the H2020 programme, aims at developing a scalable environment for data management and computing, addressing the problems of the growing data volume and focused in providing a complete framework for research communities through the European Open Science Cloud. The target of this project is to integrate different services and tools based on Cloud Computing to manage Big Data sources, and Use Cases from diverse disciplines are represented. One of the goals of the project is to deal with extremely large and heterogeneous datasets, including diverse data and metadata types, formats and standards that enable the automatic integration of Big Data.

The growing number of environmental data sources provides new possibilities to understand the complex Earth system. Remote sensing data provided by satellites, historical and forecasting data provided by meteorological agencies or in-situ sensors gives a lot of information to monitor an ecosystem. However, they are very heterogeneous in terms of file formats and access.
One of the XDC Use Cases representing LifeWatch ERIC, the European Research Infrastructure for Ecosystems and Biodiversity, is integrating data from those heterogeneous data sources for Environmental data such as Satellites (NASA Landsat, ESA Sentinel), meteorological stations (both historical and forecasting data) or In-situ instrumentation. These sources produce data in different formats like NetCDF4, HDF5 or CSV, and they are accessible via different types of APIs. The goal of this Use Case is to automatize different stages of the data lifecycle in order to simulate freshwater environments like reservoirs to forecast the hydrodynamics and water quality, facing the problem of eutrophication.

In order to interoperate those big data sources, the adoption of the four FAIR principles is needed. The use of metadata standards like the Ecological Metadata Language is the way proposed to interoperate the different data sources, and the use of Cloud Computing solutions provided by XDC the way to manage the complex data life cycle in a FAIR manner: ingestion, curation, analysis, modeling, publishing, etc.

The XDC LifeWatch ERIC Use Case proposes a Virtual Research Environment (VRE) deployed on the Cloud that allows the users to access and get different types of environmental data in an easy manner, and without the need of using local resources. In order to achieve this goal, the architecture of this VRE is designed to use the following components:

-JupyterHub as User Interface. A docker container is deployed for each user and it provides a private environment. Users are logged-in with INDIGO IAM.

-Onedata: distributed storage system to manage the ingested data and the model outputs. It supports metadata attachment and discovery, enabling metadata searching to find the data needed to feed a new model automatically.

-Orchestrator: deploys the data ingestion, analysis and different tasks with computing needs. It also enables workflows to link different tasks in the data lifecycle.

This presentation will describe the architectural design of the VRE and the different components, as well as details on how this cloud-based approach can be adopted to many other cases.