Geophysical Research Abstracts
Vol. 20, EGU2018-5328, 2018
EGU General Assembly 2018


# XDC Project: Dealing with extremely large and heterogeneous datasets

Fernando Aguilar Gómez and Jesús Marco
IFCA-CSIC, Santander, Spain (aguilarf@ifca.unican.es)

The eXtreme-DataCloud project, under the umbrella of H2020 programme, aims at developing a scalable environment for data management and computing. The target of this project is to integrate different services and tools based on Cloud Computing to manage Big Data sources, and several case studies from different disciplines (Astrophysics, Biodiversity, Environmental Sciences) 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.

One of the case studies representing LifeWatch ESFRI, the European Research Infrastructure for Ecosystems and Biodiversity, will integrate data from heterogeneous data sources for Environmental data such as Satellites (NASA Landsat, ESA Sentinel), meteorological stations, In-situ instrumentation or Internet of Things. These data sources produce data in different formats like NetCDF4, HDF5, CSV. The goal of this case study is to automatize different stages of data life cycle in order to model and simulate water environments like reservoir or lakes to forecast the hydrodynamics and water quality. In order to interoperate the big data sources, metadata standards like the Ecological Metadata Language will play a very important role to support FAIR (Findable, Accessible, Interoperable, Reusable) data production.

The poster will show the initial XDC architecture describing how solutions and services based on Cloud Computing can support the automatic management of Big and Heterogeneous Datasets and how they can contribute to orchestrate the deployment of services oriented to provide complex resources to scientists in a user-friendly manner, stimulating the development of collaborative environments. Furthermore, different technical solutions to manage big data life cycle will be presented, taking into account the heterogeneity of data and metadata formats.