Geophysical Research Abstracts
Vol. 18, EGU2016-18280, 2016
EGU General Assembly 2016

Toward server-side, high performance climate change data analytics in the Earth System Grid Federation (ESGF) eco-system

Sandro Fiore (1), Dean Williams (2), and Giovanni Aloisio (3)

(1) Fondazione Centro Euro Mediterraneo sui Cambiamenti Climatici, (2) Lawrence Livermore National Laboratory, (3) Università del Salento and Fondazione Centro Euro Mediterraneo sui Cambiamenti Climatici

In many scientific domains such as climate, data is often n-dimensional and requires tools that support specialized data types and primitives to be properly stored, accessed, analysed and visualized. Moreover, new challenges arise in large-scale scenarios and eco-systems where petabytes (PB) of data can be available and data can be distributed and/or replicated (e.g., the Earth System Grid Federation (ESGF) serving the Coupled Model Intercomparison Project, Phase 5 (CMIP5) experiment, providing access to 2.5PB of data for the Intergovernmental Panel on Climate Change (IPCC) Fifth Assessment Report (AR5).

Most of the tools currently available for scientific data analysis in the climate domain fail at large scale since they: (1) are desktop based and need the data locally; (2) are sequential, so do not benefit from available multicore/parallel machines; (3) do not provide declarative languages to express scientific data analysis tasks; (4) are domain-specific, which ties their adoption to a specific domain; and (5) do not provide a workflow support, to enable the definition of complex “experiments”.

The Ophidia project aims at facing most of the challenges highlighted above by providing a big data analytics framework for eScience. Ophidia provides declarative, server-side, and parallel data analysis, jointly with an internal storage model able to efficiently deal with multidimensional data and a hierarchical data organization to manage large data volumes (“datacubes”). The project relies on a strong background of high performance database management and OLAP systems to manage large scientific data sets. It also provides a native workflow management support, to define processing chains and workflows with tens to hundreds of data analytics operators to build real scientific use cases.

With regard to interoperability aspects, the talk will present the contribution provided both to the RDA Working Group on Array Databases, and the Earth System Grid Federation (ESGF) Compute Working Team.

Also highlighted will be the results of large scale climate model intercomparison data analysis experiments, for example: (1) defined in the context of the EU H2020 INDIGO-DataCloud project; (2) implemented in a real geographically distributed environment involving CMCC (Italy) and LLNL (US) sites; (3) exploiting Ophidia as server-side, parallel analytics engine; and (4) applied on real CMIP5 data sets available through ESGF.