Geophysical Research Abstracts
Vol. 19, EGU2017-19112, 2017
EGU General Assembly 2017

Large scale and cloud-based multi-model analytics experiments on climate change data in the Earth System Grid Federation

Sandro Fiore (1), Marcin Płóciennik (2), Charles Doutriaux (3), Ignacio Blanquer (4), Roberto Barbera (5,6), Giacinto Donvito (5), Dean N. Williams (3), Valentine Anantharaj (7), Davide D. Salomoni (5), Giovanni Aloisio (1,8)

(1) CMCC Foundation, Lecce, Italy (sandro.fiore@cmcc.it), (2) Poznan Supercomputing and Networking Center, Poland , (3) Lawrence Livermore National Laboratory, USA , (4) Universitat Politècnica de València, Spain, (5) Italian National Institute of Nuclear Physics, Italy, (6) University of Catania, Italy, (7) Oak Ridge National Laboratory, USA, (8) University of Salento, Italy

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, such as 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).
A case study on climate models intercomparison data analysis addressing several classes of multi-model experiments is being implemented in the context of the EU H2020 INDIGO-DataCloud project. Such experiments require the availability of large amount of data (multi-terabyte order) related to the output of several climate models simulations as well as the exploitation of scientific data management tools for large-scale data analytics. More specifically, the talk discusses in detail a use case on precipitation trend analysis in terms of requirements, architectural design solution, and infrastructural implementation. The experiment has been tested and validated on CMIP5 datasets, in the context of a large scale distributed testbed across EU and US involving three ESGF sites (LLNL, ORNL, and CMCC) and one central orchestrator site (PSNC).
The general “environment” of the case study relates to: (i) multi-model data analysis inter-comparison challenges; (ii) addressed on CMIP5 data; and (iii) which are made available through the IS-ENES/ESGF infrastructure.
The added value of the solution proposed in the INDIGO-DataCloud project are summarized in the following: (i) it implements a different paradigm (from client- to server-side); (ii) it intrinsically reduces data movement; (iii) it makes lightweight the end-user setup; (iv) it fosters re-usability (of data, final/intermediate products, workflows, sessions, etc.) since everything is managed on the server-side; (v) it complements, extends and interoperates with the ESGF stack; (vi) it provides a “tool” for scientists to run multi-model experiments, and finally; and (vii) it can drastically reduce the time-to-solution for these experiments from weeks to hours.
At the time the contribution is being written, the proposed testbed represents the first concrete implementation of a distributed multi-model experiment in the ESGF/CMIP context joining server-side and parallel processing, end-to-end workflow management and cloud computing.
As opposed to the current scenario based on search & discovery, data download, and client-based data analysis, the INDIGO-DataCloud architectural solution described in this contribution addresses the scientific computing & analytics requirements by providing a paradigm shift based on server-side and high performance big data frameworks jointly with two-level workflow management systems realized at the PaaS level via a cloud infrastructure.