Geophysical Research Abstracts
Vol. 18, EGU2016-18285, 2016
EGU General Assembly 2016

A big data approach for climate change indicators processing in the CLIP-C project

Alessandro D'Anca (1), Laura Conte (1), Cosimo Palazzo (1), Sandro Fiore (1), and Giovanni Aloisio (2)

(1) Fondazione Centro Euro Mediterraneo sui Cambiamenti Climatici, (2) Università del Salento and Fondazione Centro Euro Mediterraneo sui Cambiamenti Climatici

Defining and implementing processing chains with multiple (e.g. tens or hundreds of) data analytics operators can be a real challenge in many practical scientific use cases such as climate change indicators. This is usually done via scripts (e.g. bash) on the client side and requires climate scientists to take care of, implement and replicate workflow-like control logic aspects (which may be error-prone too) in their scripts, along with the expected application-level part. Moreover, the big amount of data and the strong I/O demand pose additional challenges related to the performance. In this regard, production-level tools for climate data analysis are mostly sequential and there is a lack of big data analytics solutions implementing fine-grain data parallelism or adopting stronger parallel I/O strategies, data locality, workflow optimization, etc.

High-level solutions leveraging on workflow-enabled big data analytics frameworks for eScience could help scientists in defining and implementing the workflows related to their experiments by exploiting a more declarative, efficient and powerful approach. This talk will start introducing the main needs and challenges regarding big data analytics workflow management for eScience and will then provide some insights about the implementation of some real use cases related to some climate change indicators on large datasets produced in the context of the CLIP-C project - a EU FP7 project aiming at providing access to climate information of direct relevance to a wide variety of users, from scientists to policy makers and private sector decision makers.

All the proposed use cases have been implemented exploiting the Ophidia big data analytics framework. The software stack includes an internal workflow management system, which coordinates, orchestrates, and optimises the execution of multiple scientific data analytics and visualization tasks. Real-time workflow monitoring execution is also supported through a graphical user interface. In order to address the challenges of the use cases, the implemented data analytics workflows include parallel data analysis, metadata management, virtual file system tasks, maps generation, rolling of datasets, and import/export of datasets in NetCDF format.

The use cases have been implemented on a HPC cluster of 8-nodes (16-cores/node) of the Athena Cluster available at the CMCC Supercomputing Centre. Benchmark results will be also presented during the talk.