Geophysical Research Abstracts
Vol. 17, EGU2015-15101-2, 2015
EGU General Assembly 2015

Big data analytics workflow management for eScience

Sandro Fiore, Alessandro D'Anca, Cosimo Palazzo, Donatello Elia, Andrea Mariello, Paola Nassisi, and Giovanni Aloisio

Euro Mediterranean Center on Climate Change (CMCC)

In many domains such as climate and astrophysics, scientific data is often n-dimensional and requires tools that support specialized data types and primitives if it is to be properly stored, accessed, analysed and visualized. Currently, scientific data analytics relies on domain-specific software and libraries providing a huge set of operators and functionalities. However, most of these software fail at large scale since they:
(i) are desktop based, rely on local computing capabilities and need the data locally;
(ii) cannot benefit from available multicore/parallel machines since they are based on sequential codes;
(iii) do not provide declarative languages to express scientific data analysis tasks, and
(iv) do not provide newer or more scalable storage models to better support the data multidimensionality.
Additionally, most of them:
(v) are domain-specific, which also means they support a limited set of data formats, and
(vi) do not provide a workflow support, to enable the construction, execution and monitoring of more 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 several parallel operators to manipulate large datasets. Some relevant examples include: (i) data sub-setting (slicing and dicing), (ii) data aggregation, (iii) array-based primitives (the same operator applies to all the implemented UDF extensions), (iv) data cube duplication, (v) data cube pivoting, (vi) NetCDF-import and export. Metadata operators are available too. Additionally, the Ophidia framework provides array-based primitives to perform data sub-setting, data aggregation (i.e. max, min, avg), array concatenation, algebraic expressions and predicate evaluation on large arrays of scientific data. Bit-oriented plugins have also been implemented to manage binary data cubes.
Defining processing chains and workflows with tens, hundreds of data analytics operators is the real challenge in many practical scientific use cases. This talk will specifically address the main needs, requirements and challenges regarding data analytics workflow management applied to large scientific datasets. Three real use cases concerning analytics workflows for sea situational awareness, fire danger prevention, climate change and biodiversity will be discussed in detail.