Contribution ID: **4** Type: **not specified**

The Ophidia framework: toward cloud-based big data analytics for eScience

Thursday, 25 September 2014 11:50 (20 minutes)

In many domains such as life sciences, 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, analyzed and visualized. The n-dimensionality of scientific datasets, and their data cube abstraction, leads to a need for On-Line Analytical Processing (OLAP)-like primitives such as slicing, dicing, pivoting, drill-down, and roll-up. These primitives have long been supported in data warehouse systems and used to perform complex data analysis, mining and visualization tasks. Unfortunately, current OLAP systems fail at large scale—different storage models and data management strategies are needed to fully address scalability. Yet the analysis of scientific datasets has a higher computing demand with regard to current OLAP systems, which definitely leads to the need of having parallel/distributed solutions to meet the (near) real-time requirement. Finally OLAP systems are domain agnostic, so they do not provide domain-based support, functions and primitives that are essential to fully address scientific analysis.

Currently, scientific data analytics relies on domain-specific software and libraries providing a huge set of operators and functionalities. This approach will fail at the large scale, because most of these software: (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.

A related work in this area is the Ophidia project, a research effort on big data analytics facing scientific data analysis challenges in the climate change domain. It provides parallel (server-side) data analysis, an internal storage model and a hierarchical data organization to manage large amount of multidimensional scientific data. The Ophidia analytics platform provides several MPI-based parallel operators to manipulate (as a whole) the entire set of fragments associated to a data cube. 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. 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. Multiple primitives can be nested to implement a single more complex task (e.g., aggregating by sum a subset of the entire array). Bit-oriented plugins have also been implemented to manage binary data cubes; compression algorithms can be included as primitives too.

The entire Ophidia software stack has been deployed at CMCC on 24-nodes (16-cores/node) of the Athena HPC cluster. A comprehensive benchmark and test cases are being defined with climate scientists to extensively test all of the features provided by the system. Preliminary experimental results are already available and have been published on scientific research papers.

The most relevant data analytics use cases implemented in national and international projects target fire danger prevention (OFIDIA), sea situational awareness (TESSA), interactions between climate change and biodiversity (EUBrazilCC), climate indicators and remote data analysis (CLIP-C), large scale data analytics on CMIP5 data in NetCDF format, Climate and Forecast (CF) convention compliant (ExArch).

In particular, in the context of the EU FP7 EUBrazil Cloud Connect project (http://eubrazilcloudconnect.eu/), the Ophidia framework is being extended in order to integrate scalable VM-based solutions for the management of large volumes of scientific data (both climate and satellite data) in a cloud-based environment to study how climate change affects biodiversity.

Primary author: Dr FIORE, Sandro (CMCC)

Co-authors: Dr WILLIAMS, Dean (LLNL); Prof. ALOISIO, Giovanni (CMCC and University of Salento); Prof. FOSTER, Ian (ANL and Univ. of Chicago)

Presenter: Dr FIORE, Sandro (CMCC)

Session Classification: EGI-GEANT Symposium: Platforms