Contribution ID: **50** Type: **Presentation**

The Ophidia stack: a big data analytics framework for Virtual Research Environments

Wednesday, 11 November 2015 14:30 (20 minutes)

The Ophidia project is a research effort on big data analytics facing scientific data analysis challenges in multiple domains (e.g. climate change). It provides a framework responsible for atomically processing and manipulating datacubes, by providing a common way to run distributive tasks on large set of fragments (chunks).

Even though the most relevant use cases for Ophidia have been implemented in the climate change context, the domain-agnostic design of the internal storage model, operators and primitives makes easier the exploitation of the framework as a core big data technology for multiple Research Communities.

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. The project relies on a strong background on high performance database management and OLAP systems to manage large scientific datasets.

The Ophidia analytics platform provides several data operators to manipulate data cubes, and array-based primitives to perform data analysis on large scientific data arrays (e.g. statistical analysis, predicate evaluation, FFT, DWT, subsetting, aggregation, compression). The array-based primitives are built on top of well-known numerical libraries (e.g. GSL). Bit-oriented primitives are also available to manage B-cubes (binary data cubes). Metadata management support (CRUD-like operators) is also provided jointly with validation-based features relying on community/project-based vocabularies.

The framework 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. Defining processing chains and workflows with tens, hundreds of data analytics operators can be a real challenge in many practical scientific use cases. The talk will also highlight the main needs, requirements and challenges regarding data analytics workflow management applied to large scientific datasets.

Some real use cases implemented at the Euro Mediterranean Center on Climate Change (CMCC) will be also discussed. The results of a benchmark performed on the Athena Cluster at the CMCC SuperComputing Centre and regarding CMIP5 datasets will be also presented.

Primary author: Dr FIORE, Sandro (CMCC)

Co-author: Prof. ALOISIO, Giovanni (CMCC & University of Salento)

Presenter: Dr FIORE, Sandro (CMCC)

Session Classification: Exploiting the EGI Federated clouds - Paas & SaaS workshop