Contribution ID: **88** Type: **Sessions contributions**

Big data analytics in the EUBrazil Cloud Connect project

Wednesday, 21 May 2014 16:00 (20 minutes)

The EUBrazil Cloud Connect (EUBrazilCC) project is a first step towards providing a user-centric, cross-Atlantic test bench for European & Brazilian research communities. It includes the implementation of three multidisciplinary & highly complementary scenarios, covering Epidemiology, Medical Systems simulation, Biodiversity & Climate Change.This contribution is strongly related to the 3rd use case which has a special focus on big data analytics and it is built on a close collaboration among European & Brazilian excellence centres. It deals with multiple multidimensional and heterogeneous data sources like output of global and regional climate models simulations and different types of satellite data. To address the use case requirements, the EUBrazilCC project is providing a cloud framework for big data analytics (named PDAS) joining novel storage models, HPC and parallel database management solutions. The framework carries out (near) real time data analytics tasks (e.g. data reduction, time series analysis, data slicing) on large multidimensional datasets, exploiting parallel data operators (MPI/OpenMP based). The PDAS exposes a WS-I interface (GSI/VOMS enabled) to interoperate with the EGI infrastructure. The implementation of a cloud interface to enable elastic resource provisioning as well as the interaction with the COMPSs framework to provide workflow-based analytics on massive volumes of data are major goals to be addressed during the EUBrazilCC project implementation.

Wider impact and conclusions

The workflows commonly used for scientific discovery (based on search, locate, download & analyze steps) will fail at large scale due to time- and resource-consuming data downloads and need of big computing facilities. The PDAS exploits a different approach based on server-side analysis capabilities exploiting data-intensive facilities close to data storage. It reduces the downloaded data (e.g. maps, summaries), the makespan for the analysis task, and the complexity for the software requirements on client machines.
COMPSs is already adopted in the EGI Federated Cloud to provide scalability and elasticity capabilities for the deployment of biodiversity workflows. The adoption of COMPSs to implement complex workflows on top of the PDAS service will benefit the project's use cases through the optimized execution of sequences of data analytics operations. PDAS will be used to analyse massive retrospective Brazilian climate and vegetation data in the frame of EUBrazilCC.

URL(s) for further info

www.eubrazilcloudconnect.eu (under construction)

Description of work

One of the tasks of the EUBrazilCC project is the implementation of a big data analytics framework (PDAS) exploiting scalable VM-based solutions for the management of large volumes of scientific data (e.g. output of climate simulations and satellite data). The PDAS leverages HPC paradigms (MPI/OpenMP based) and database technologies (both relational and key-value based). The system exposes a GSI and VOMS enabled WS-I interface.To address efficient and flexible analysis of multidimensional datasets, the PDAS implements a novel storage model jointly with a hierarchical storage management. The datasets stored into the system are referenced through DOIs.The framework provides a set of primitives to manipulate the data, perform time series analysis, data sub-setting, data reduction, etc. A wide set of parallel "datacube" oriented functionalities (e.g. reduce, subset, merge, split) are already implemented and available to the end users. A terminal-like client application provides a useful set of commands to remotely submit analytics operators. To ease the integration of the PDAS service with the other components of the EUBrazilCC infrastructure and with the use cases application, the project will provide workflow functionalities through the COMPSs framework. COMPSs is able to orchestrate the execution of the tasks composing an application, both regular methods or web services,

taking care of the data dependencies between the invocations. The implemented COMPSs-PDAS application will be offered as additional service in the project infrastructure, allowing to run DAGs of multiple operators which enact relevant processing chains for the Biodiversity & Climate Change use case. A key challenge of the PDAS is related to the management of metadata which is addressed at two different levels: system and application. In this regard a complete set of operators provide CRUD-like metadata support, search & discovery, vocabulary definition and provenance management.

Primary author: Dr FIORE, Sandro (Euro Mediterranean Center on Climate Change (CMCC))

Co-authors: LEZZI, Daniele (Barcelona Supercomputing Center); Prof. ALOISIO, Giovanni (University of Salento and CMCC); Dr BLANQUER, Ignacio (UPVLC); Dr BADIA, Rosa (BSC)

Presenter: Dr FIORE, Sandro (Euro Mediterranean Center on Climate Change (CMCC))

Session Classification: New data management solutions for EGI

Track Classification: Requirements and solutions for data management and computing (Track Leaders: B. Konya, H. Heller, S. Tarkoma)