Cloudscape IV – Advances on Interoperability & Cloud Computing Standards
Dates: 23rd & 24th February 2012
Location: Avenue de Beaulieu BU #25, Brussels Belgium (Ground Floor, Nearest Metro Beaulieu)

PRESENTATION TITLE: Powering Science by embedding Cloud Computing in Hybrid Data Infrastructures

PRESENTER NAME: Pasquale Pagano

Abstract

A recent study promoted by The Royal Society analysed the changing patterns of science and scientific collaborations and confirmed that science is increasingly global, multipolar, and networked. This trend is captured by the VVV paradigm, that is, data growth in *Volume*, *Variety*, and collection, processing and consumption *Velocity.*

The requirements of this new science outgrow the capacity of traditional approaches. In particular:

- Global, multipolar and networked scientific collaborations require dynamic computational environments capable of dealing with the VVV paradigm.

- Heterogeneity of the data types and data sources to be integrated requires computational environments capable of federating a number of different technologies.

- Innovative solutions for massive data storage, curation, management, and analysis require elastic access and usage of computational resources.

The ***Hybrid Data Infrastructure*** (HDI) is an emerging paradigm, which assumes that different technologies, including cloud computing can be integrated to provide elastic access and usage of data and data-management capabilities needed to address the challenges of new science.

The iMarine[1] and EUBrazilOpenBio[2] projects are underpinned by Hybrid Data Infrastructures. These infrastructures support world-wide collaboration between distributed biodiversity communities by federating heterogeneous data sources, offering innovative solutions to scientists. This is achieved by integrating advanced technologies for the management of data and by providing access to heterogeneous cloud computing resources by embedding the VENUS-C[3] technology.

Relevance to the 3 pillars of the European Commission's Cloud Strategy

The availability of computational and storage resources offered by commercial cloud providers, the value-add of the new technologies generally identified as no-sql databases, the accessibility of distributed computing platform supporting MapReduce[4], porting to MapReduce of several algorithms for performing data analysis and mining, and the open-access movement ensures increasing availability of data sources are all completely changing the support that ICT offers to scientists.. All these opportunities can be leveraged thanks to the innovative vision introduced by the ***Hybrid Data Infrastructure:*** integrating technologies to support efficient data management.

Our work places special emphasis on the design and operation of HDIs, aiming to promote innovation in science by ensuring data interoperability and efficient data management in a secure

[1] http://www.i-marine.eu
[2] http://www.eubrazilopenbio.org
[3] http://www.venus-c.eu
[4] http://www.mapreduce.org/

yet dynamic environment where the provider of the HDI ensures data privacy, governance, and attribution.

Field of application

Hybrid Data Infrastructures offer a solution to any global, multipolar, and networked scientific community. They offer the power of cloud computing enriched with the accessibility of heterogeneous data sources maintained by third-parties. They enable scalability and cost-effectiveness. They enable the efficient exchange of scientific discovery through access to the most advanced technologies while hiding the complexity of computational platforms..

iMarine and EUBrazilOpenBio are both serving diverse scientific communities that are facing new challenges, such as biologists generating model-based large-scale predictions of the natural occurrences of species, statisticians managing and integrating statistical data, and biodiversity scientists willing to build, test and project models of species distribution.

Standards and Interoperability

Hybrid Data Infrastructures aim to integrate resources potentially from multiple from cloud providers. This approach is still costly due to the absence of standards for the management of cloud resources and exploitation of such resources.

Examples of possible standards at IaaS management level are:

- standards for describing resource/performance capabilities and requirements
- standards for monitoring, auditing, billing, reporting, and notification
- standards for policies and governance
- standards for federated security across Cloud infrastructures
- standards for machine readable SLAs (Service Level Agreements).

The projects leveraging the new infrastructures are following activities related to CDMI (SNIA) and OCCI (OGF) but are not directly involved in any standard-related group.

Standards enabling interoperability and portability of applications across clouds are needed to avoid vendor lock-in and allow the development of advanced services built on top of cloud infrastructures.

The following use cases need to be prioritised:

- federated security across clouds
- moving applications across clouds
- moving data and metadata across clouds
- running application workflows across clouds

Cloud computing technologies and cloud infrastructures are valuable in the emerging landscape. Interoperability across clouds, even if achieved to a certain degree, will boost research across a spectrum of scientific communities.

Contact Person **Pasquale Pagano**

Web: http://it.linkedin.com/in/pasqualepagano

Contact details: pasquale.pagano@isti.cnr.it -- CNR – ISTI, Via G. Moruzzi 1, 56124 Pisa – Italy.

Relevant Links: www.gcube-system.org, www.i-marine.eu, www.eubrazilopenbio.org, http://www.venus-c.eu