



Contribution ID: 40

Type: **Oral Presentation**

e-Infrastructures Integration with gCube

Wednesday, 13 April 2011 15:00 (30 minutes)

Overview

Delivering an e-Infrastructure service to large organizations is a complex task that requires the integration of several technologies. The complexity of this service resides on: (i) very rich applications and data collections are maintained by a multitude of authoritative providers; (ii) different problems require different execution paradigms: batch, map-reduce, synchronous call, message-queue, etc.; (iii) key distributed computation middleware exist: gLite, Globus, and Unicore for grid-based wide resource sharing; Condor for site resource sharing; Hadoop and Cassandra for cluster resource sharing; etc.; (iv) several standards exist in the same domain.

gCube, the D4Science empowering technology, offers solutions to abstract over differences in location, protocols, and models by scaling no less than the interfaced resources, by keeping failures partial and temporary, and by being autonomic reacting and recovering from a large number of potential issues.

Impact

gCube doesn't hide the infrastructure middleware. It is not another layer. Rather it turns infrastructures into a utility by offering a single submission, monitoring, and access facilities. It offers a common framework to programming in the large and in the small. It allows to exploit concurrently private virtualized resources organized in sites with resources provided by IaaS and PaaS cloud providers.

By using its set of facilities several scientific applications have been implemented and delivered to the Fishery and Aquaculture Resource Management communities delivering the following applications (among others):

- A collaboration-oriented suite providing seamless access and organisation facilities on a rich array of objects (e.g. information objects, queries, files, templates, timeseries). It offers mediation capabilities between the external world objects, systems and infrastructures (import/export/publishing) and it supports common file management features (drag & drop, contextual menu);
- A timeseries framework offering tools to manage large datasets by supporting the complete timeseries lifecycle (validation, curation, analysis, and reallocation). It offers tools to operate on multi-dimensional statistical data. It supports filtering, grouping, aggregation, union, mining, and plotting;
- An ecological niche modeling suite to predict the global distribution of marine species, that generates color-coded species range maps using half-degree latitude and longitude blocks by interfacing several scientific species-databases and repository providers. It allows the extrapolation of known species occurrences to determine environmental envelopes (species tolerances) and to predict future distributions by matching species tolerances against local environmental conditions (e.g. climate change and sea pollution).

Description of the work

gCube is a large software framework designed to abstract over several technologies and offer them through a well-formed set of APIs. gCube consists of several packages offering:

- Access to several storage back-ends tailored for different needs. For example, it offers a storage server: for multiple-version software packages; for scientific data-sets stored as tables; for timeseries with an OLAP interface; for structured document objects; for geo-coded datasets compliant with OGC; and finally for storing files;
 - Management of metadata in any format and schema that can be consumed by the same application in the same Virtual Organization;
- A process execution engine to manage the execution of software elements in a distributed infrastructure under the coordination of a composite plan that defines the data dependencies among its actors. It supports several computational middleware without performance compromises. A task can be designed as a workflow of invocation of different code components (services, binary executables, scripts, map-reduce jobs, etc.);
- A transformation engine to transform data among various manifestations. This engine is manifestation and transformation agnostic by offering an object-driven operation workflow. It is extensible through the addition of transformation-program plugins;
 - A Virtual Research Environment (VRE). Through VREs, groups of users have controlled access to distributed data, services, storage, and computational resources integrated under a personalised environment. VREs support cooperative activities such as: metadata cleaning, enrichment, and transformation by exploiting mapping schema, controlled vocabulary, thesauri, and ontology; processes refinement and show cases implementation; data assessment; expert users validation of products generated through data elaboration or simulation; sharing of data and process with other users.

URL

<http://www.gcube-system.org/>

Conclusions

gCube is currently deployed in the D4Science production infrastructure and provides to the Fishery and Aquaculture Resource Management communities a large number of APIs and highly specialized scientific applications running on an e-Infrastructure service which hides the complexity of the underlying heterogeneous set of middleware systems, standards, data types, metadata schemas, etc.

Primary authors: MANZI, Andrea (CERN); PAGANO, Pasquale (CNR); ANDRADE, Pedro (CERN)

Presenters: MANZI, Andrea (CERN); PAGANO, Pasquale (CNR); ANDRADE, Pedro (CERN)

Session Classification: User Environments

Track Classification: User Environments - Applications