Contribution ID: **12** Type: **Presentation**

DARE: Integrating solutions for Data-Intensive and Reproducible Science

Wednesday, 8 May 2019 11:30 (15 minutes)

The DARE (Delivering Agile Research Excellence on European e-Infrastructures) project is implementing solutions to enable user-driven reproducible computations that involve complex and data-intensive methods. The project aims at providing a platform and a working environment for professionals such as Domain Experts, Computational Scientists and Research Developers. These can compose, use and validate methods that are expressed in abstract terms. DARE's technology translates the scientists'workflows to concrete applications that are deployed and executed on cloud resources offered by European e-infrastructures, as well as in-house institutional platforms and commercial providers. The platforms'core services enable researchers to visualise the collected provenance data from runs of their methods for detailed diagnostics and validation. This also helps them manage long-running campaigns by reviewing results from multiple runs. They can exploit heterogeneous data effectively.

To shape and evaluate the integrated solutions, the project analysed a variety of demanding use cases. Prototypes and agile co-development ensure solutions are relevant and reveal the priority issues. Our use cases are presented by two scientific communities in the framework of EPOS and IS-ENES, conducting respectively research in computational seismology and climate-impact studies. We will present how DARE enables users to develop and validate their methods within generic environments such as Jupyter notebooks associated with conceptual and evolving workspaces, or via the invocation of OGC WPS services. These different access modes will be integrated in the architecture of the systems already developed by the two target communities, interfacing with institutional data archives, incrementally accommodating complex computational scenarios and reusable workflows.

We will show how DARE exploits computational facilities adopting software containerisation and infrastructure orchestration technologies (Kubernetes). These are transparently managed from the DARE API, in combination with registries describing data, data-sources and methods. Ultimately, the extensive adoption of workflows (dispel4py, CWL), methods abstraction and containerisation, allows DARE to dedicate special attention to portability and reproducibility of the scientists'progress in different computational contexts. This enables the optimised mapping to new target-platform choices without requiring method alteration and preserving semantics, so that methods continue to do what users expect, but they do it faster or at less cost. Methods can avoid unnecessary data movement and combine steps that require different computational contexts.

Validation and monitoring services are implemented on top of a provenance management system that combines and captures interactive and lineage patterns, These are expressed in W3C PROV compliant formats and represent research developers'choices, the effects on the computational environment and the relationship between data, processes and resources, when their workflows are being executed. We will discuss how the patterns are modelled and implemented using Templates and Lineage services (Provenance Template Registry, S-ProvFlow), in order to integrate and interactively use context-rich provenance information.

Type of abstract

Presentation

References

http://project-dare.eu

Primary author: SPINUSO, Alessandro (KNMI)

Co-authors: KRAUSE, Amy (EPCC); GEMUEND, Andre (Fraunhofer); KLAMPANOS, Iraklis (National Centre for Scientific Research "Demokritos"); ATKINSON, Malcolm (UE); FILGUEIRA, Rosa (EPCC)

Presenter: SPINUSO, Alessandro (KNMI)

Session Classification: Scientific and technical updates