Enabling Transparent Access to Heterogeneous Architectures for IS-ENES climate4impact using the DARE Platform

Christian Pagé
CECI
Université de Toulouse, CNRS, Cerfacs
Toulouse, France
christian.page@cerfacs.fr

Wim Som de Cerff
KNMI
Koninklijk Nederlands Meteorologisch Instituut
De Bilt, Netherlands
wim.som.de.cerff@knmi.nl

Maarten Plieger
KNMI
Koninklijk Nederlands Meteorologisch Instituut
De Bilt, Netherlands
maarten.plieger@knmi.nl

Alessandro Spinuso
KNMI
Koninklijk Nederlands Meteorologisch Instituut
De Bilt, Netherlands
alessandro.spinuso@knmi.nl

Xavier Pivan
Koninklijk Nederlands Meteorologisch Instituut
De Bilt, Netherlands
xavier.pivan.ds@gmail.com

Abstract—Access to Climate data is crucial to sustain research and climate change impact assessments. It has a strong societal impact as those changes will have to be mitigated as much as possible. The whole climate data archive is expected to reach a volume of 30 PB in 2019 and up to 2000 PB in 2024 (estimated), evolving from 30 TB in 2007 and 2 PB in 2014. Data processing and analysis must now happen remotely for the users: they now have to rely on heterogeneous infrastructures and services between the data and their location. Developers of Research Infrastructures have to provide services to those users, hence having to define standards and generic services to fulfill those requirements. It will be shown how the DARE eScience Platform (http://project-dare.eu) will help developers to develop more rapidly needed services for a large range of scientific researchers. The platform is designed for efficient and traceable development of complex experiments and domain-specific services on the Cloud. It will be also shown how the integration of the DARE platform together with the climate IS-ENES (https://is.enes.org) consortium has developed a platform to ease access to climate data for the climate impact community (C4I: https://climate4impact.eu). The platform is based on existing standards (ISO/OGC), such as WPS (Web Processing Service). The DARE Platform integrates services from the EUDAT CDI, enabling generic access and cross-domain interoperability, as well as providing compliance and integration with the EOSC platform. The DARE platform uses containerization technologies, so that it can be easily deployed on heterogeneous architectures. A scientific pilot has been designed within the DARE project for the ENES community (climate domain) and a prototype has been evaluated by C4I developers. It will enable delegation of on-demand computational-intensive calculations using the DARE platform, from the IS-ENES C4I interface, seamlessly and transparently to the users, especially in the context of new developments taking place within the EU H2020 IS-ENES3 project.

I. INTRODUCTION

End Users of Climate data have nowadays to struggle with accessing the data they need for their research because of the rapid increase in data volumes. The whole climate data archive is expected to reach a volume of 30 Pb late in 2019 and up to 2000 Pb in 2022 (estimated). On-demand data processing solutions as close as possible to the data storage are emerging, thanks to newly developed standards, provenance and infrastructures. In Europe several initiatives are taking place to support scientific on-demand data analytics at the European scale. They offer the huge potential of interoperability, as for example the DARE e-science platform (http://project-dare.eu/), designed for efficient and traceable development of complex experiments and domain-specific services on the Cloud. Also, the IS-ENES (https://is.enes.org) consortium has developed a platform to ease access to climate data for the climate impact community (C4I: https://climate4impact.eu). The platform is based on existing standards (ISO/OGC), such as WPS (Web Processing Service). The DARE Platform integrates services from the EUDAT CDI, enabling generic access and cross-domain interoperability, as well as providing compliance and integration with the EOSC platform. The DARE platform uses containerization technologies, so that it can be easily deployed on heterogeneous architectures. A scientific pilot has been designed within the DARE project for the ENES community (climate domain) and a prototype has been evaluated by C4I developers. It will enable delegation of on-demand computational-intensive calculations using the DARE platform, from the IS-ENES C4I interface, seamlessly and transparently to the users, especially in the context of new developments taking place within the EU H2020 IS-ENES3 project.

II. MOTIVATIONS: THE IS-ENES C4I PLATFORM AND ITS DIVERSE USERS

The IS-ENES C4I platform is being developed since 2009, currently within the IS-ENES3 H2020 EU project. It has evolved according to user consultations throughout the years as well as suggestions and feedback from users and scientific conferences [1]. Initially it was targeted to all users working in climate change impact assessments, including stakeholders.
After users requirements gathering it became obvious that the users were rather of the following categories:

1) Climate change impact modellers/researchers using data as input to impact models, such as in hydrology, crop modelling, land use, etc.
2) Expert technical offices and consultants, with technical and scientific expertise and knowledge;
3) Climate researchers themselves, especially PhD students and post-docs.

Providing a suitable web portal to support this range of users requires a very good knowledge of their needs and their own expertise and knowledge. It also requires an excellent knowledge about technologies, not only current ones but also both deprecated and emerging ones. C4I evolved from a web portal approach, providing easier data access with guidance, help, documentation, use case examples, to a full platform providing an interface along with services and APIs. Those services include on-demand data processing and statistical downscaling. One key aspect is the provision of reusable services using standard interfaces and APIs. They can be used in user tools as well as building blocks to build other thematic and/or regional and national portals.

New challenges

Because of the rapid increase in data volumes of climate data, online data processing and analyses closer to the storage is now almost unavoidable. This is especially true for end-users working in climate change impact research with limited network bandwidth, storage and computing resources. This is even getting true also for climate researchers themselves, who have access to large compute resources and storage with very fast network bandwidth. In the last inter-comparison experiment, there are high resolution climate modelling configurations along with a large increase in the number of experiments (for specific scientific questions), scenarios (for uncertainty quantification), and modelling centers.

III. INTEGRATION AND BENEFITS

The C4I platform aims at easing access to climate data, notably by providing on-demand online processing services. However it also means that the platform must be able to delegate large calculations on servers closer to the data with enough computing resources. The heterogeneity of the users means that they have access to diverse computing platforms, such as private and public clouds, national clusters, etc. Developing the C4I backend to support those diverse infra and e-infrastructures is a challenge. The DARE platform objective is to provide efficient and transparent access to diverse computing resources using an API with flexible components that can be deployed easily. By implementing the DARE platform as a backend of the C4I platform unlocks the potential of easily deploying the calculations onto heterogeneous computing platforms, taking into advantage the available parallel processing of computing resources. It also speeds up considerably the work of the C4I developers by providing them a ready to roll solution to rapidly integrate diverse e-infrastructures such as the EOSC and the EUDAT CDI B2 Services, but also onto private Clouds or custom clusters. An evaluation of the DARE platform by C4I developers has taken place on June 17th, 2019 in the Netherlands (results of this evaluation are summarized in section IV).

The DARE platform also provides automated provenance generation, using S-ProvFlow ([2]) through dispel4py ([3] and [4]) https://github.com/dispel4py/dispel4py, which is a very important aspect when delegating calculations onto one or several backends. Researchers will need quite detailed information on every computation steps for reproducibility reasons, as well as for full lineage information.

The deployment of DARE components on behalf of the user will be triggered by a C4I WPS, using the DARE API and its registry. The C4I users will use the C4I front-end (in the future a wizard) to trigger the execution of the data processing, seamlessly and transparently. It the final implementation, it will also propagate the authentication and authorization of the user where needed, such as credentials to access data and/or computing resources and eventually storage.

IV. ARCHITECTURE OF THE PROTOTYPE DESIGN

From those requirements, a first prototype has been designed (see 1). It was also evaluated on June 17th, 2019. In this prototype, the implementation is using the following components:

1) IS-ENES CDI C4I;
2) EUDAT CDI Service B2DROP;
3) ESGF RI Data Nodes;
4) DARE Components: API, Registry, dispel4py, lineage and provenance system;
5) No authentication system is yet in place in this first prototype, but placeholders are there.

This first prototype is implementing the most important interfaces needed for the full generic use case (workflow details in [5]). The most significant part being the integration of the DARE Platform API. Using this API, the workflow can use DARE components that are deployed into the DARE kubernetes container testbed: testbed.project-dare.eu. This testbed is
used for the development and evaluation phases. The DARE API is also enabling seamless access to the EUDAT CDI B2DROP Service, and is used for input and output in this prototype.

The integration with the Climate Research Infrastructure (RI) is not completed yet as this is a prototype. The C4I input form for the execution parameters is still rudimentary. Also, the input files are not yet accessed through the ESGF Data Nodes, but are rather prepared in advance and made accessible through the EUDAT B2DROP Service. Finally, the output file is not transferred yet back to the C4I user space, but is only uploaded to B2DROP.

But for the evaluation, since it is targeting the software developers of the climate RI and especially C4I, the focus for this first prototype was on the WPS itself with the integration of the DARE components. This is the main part software developers will have to deal directly with, when building the data processing services for the climate RI. This is where the DARE platform will be the most visible.

In this prototype, the WPS itself is executed within C4I, so the interface between the climate RI C4I and the DARE components take place within the WPS. This interface enables the deployment of the calculations away from the C4I server.

Evaluation Training Event

The evaluation training event was by invitation only, and some people involved in the DARE project were also present but did not fill the survey form, except for writing suggestions. The total number of attendees was 16, with 4 people remotely, and all of them were software and infrastructure developers.

All the results will not be shown here (for that refer to [5]), but the results that summarizes well are shown in figures 2 and 3. It is quite re-assuring to see the result that most of the software developers present do think that the DARE Platform Infrastructure approach is quite promising. The second very interesting result is that a large majority of the attendees think that the DARE Platform could significantly change the way the develop workflows for the climate infrastructure. This is one of the main objective of DARE and it means that the evaluation was quite convincing with this respect, even with only a demo of the prototype implementation.

V. SUMMARY AND NEXT ACTIONS

Up to now the overall feedback is quite positive, and the evaluation helped to identify better what should be the focus for the next development phases. One of the conclusions is that the interest is very high as DARE is really filling a needed gap in the infrastructures services. If suggestions and feedback are taken into account with proper dissemination and a future second evaluation, the DARE platform will be fully integrated and use within the Climate Research Infrastructure. This will also help to disseminate to other scientific domains that have similar requirements.

VI. ABBREVIATIONS AND ACRONYMS

- API: Application Programming Interface
- C4I: IS-ENES climate4impact
- DARE: Delivering Agile Research Excellence EU-funded H2020 Project
- ESGF RI: Earth System Grid Federation Research Infrastructure
- ENES: European Network for Earth System modelling
- ENES CDI: ENES Climate Data Infrastructure
- EOSC: European Open Science Cloud
- EU H2020: European Commission Horizon 2020 Research and Innovation programme
- EUDAT CDI: EUDAT Collaborative Data Infrastructure
- EUDAT CDI B2 Services: A EUDAT CDI collection of integrated services
- IS-ENES: Infrastructure of the European Network for Earth System modelling
- ISO: International Organization for Standardization
- OGC: Open Geospatial Consortium
- WPS: Web Processing Service

ACKNOWLEDGEMENTS

DARE builds on a succession of prior EU projects: ADMIRE, VERCE, ENVRI, ENVRIplus, and on nationally funded projects. It depends on the coordination of e-Infrastructures by EOSC. The authors also thank the developers and systems teams that have delivered and tested the platform. IS-ENES3 is the main project driving and developing the climate infrastructure in Europe, fully funded by EU. It has been preceded with...
two past EU projects IS-ENES and IS-ENES2. The authors thank especially all people involved in the development of C4I, especially Maarten and Wim working at KNMI. The DARE project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 777413 while the IS-ENES3 project is under grant agreement No 824084.

REFERENCES

[1] C. Pagé, M. Vega, A. Cofino, C. Jack, M. Plieger., Review CLIMATE4IMPACT services and objectives, EC FP7-IS-ENES2 Project Milestone (2016).

[2] Provenance framework for storage and access of data-intensive streaming lineage, https://github.com/KNMI/s-provenance, accessed: April 2019.

[3] R. Filgueira, A. Krause, M. Atkinson, I. Klampanos, A. Moreno, dispel4py: A Python Framework for Data-Intensive Scientific Computing, International Journal of High Performance Computing Applications.

[4] R. Filgueira, A. Krause, M. Atkinson, I. Klampanos, A. Spinusso, S. Sanchez-Exposito, dispel4py: An agile framework for data-intensive esicence, in: Proc. IEEE e-Science ’15, 2015, pp. 454–464. doi:10.1109/eScience.2015.40.

[5] C. Pagé., Pilot Tools and Services, Execution and Evaluation Report I, EC H2020-DARE Project Deliverable (2019).