Geophysical Research Abstracts
Vol. 16, EGU2014-2689, 2014
EGU General Assembly 2014

EUDAT and ENES: Big Climate Data Analysis

Stephan Kindermann and Hannes Thiemann
German Climate Computing Centre (DKRZ), Hamburg, Germany (kindermann@dkrz.de)

The European Network for Earth System modeling (ENES) community is facing an unprecedented requirement to provide data analysis facilities at data centers and integrate these with end-user data analysis workflows.
The existing distributed ENES data infrastructure integrates European data centers and is part of the international Earth System Grid Data Federation which stores Petabytes of climate modeling and observation data and provides this data via consistent interfaces to end users. The P2P architecture of the ESGF infrastructure provides flexible and scalable data search and data access services up to now mostly used in "download and process at home" data analysis scenarios.

We will provide an overview of the approaches taken to enable future distributed scalable data analysis workflows interfacing the ENES data federation. Multiple parallel activities are ongoing and necessary to address the different Big Data Analysis requirements:
- Several data centers have started the development of OGC Web Processing Service (WPS) frameworks to expose data analysis functionality to end users via standardized, self-describing web interfaces. The European IS-ENES project as well as the G8 exascale project ExArch are supporting this effort.
- A stable and fast data replication infrastructure between data centers and between computing and data centers is necessary to support data preservation and data collection activities. ENES is partner in the EUDAT project developing the basic infrastructure services.
- ENES and ESGF are starting collaboration and research projects with the international data network providers to improve bulk data transport activities.
- First efforts are taken to evaluate existing scalable big data storage and analysis systems (e.g. based on hadoop, openstack). By now specific "at site" deployments seem to be preferable to outsourcing solutions to commecial cloud infrastructure providers.

Besides the technical implications of future ENES big data analysis activities on the ENES data infrastructure we will also shortly discuss the non-technical data analysis requirements induced by the future multi-community usage scenarios of climate data, requiring e.g. more detailed metadata, data documentation and data provenance services.