Geophysical Research Abstracts
Vol. 13, EGU2011-10324-1, 2011
EGU General Assembly 2011

Infrastructure for Data-Intensive Seismology: Cross-correlation of distributed seismic traces through the ADMIRE framework

Alessandro Spinuso, Luca Trani, Malcolm Atkinson, and Michelle Galea
School of Informatics, University of Edinburgh, United Kingdom (school-office@inf.ed.ac.uk / Fax: +44 (0) 131 651 1426)

The objective of the presented work is the design and the implementation of a prototypal Infrastructure for data-intensive seismology, adopting the new technical solution provided by the ADMIRE (Advanced Data Mining and Integration Research for Europe, http://www.admire-project.eu) framework.

During an intense six months experience we have developed a deep understating of the ADMIRE technology, evaluating its potential in tackling the most important needs in the field of the cross-correlation of ambient seismic noise, which exploits large and continuously-recorded datasets (In the form of time series at discrete spatial points) to answer questions related to the property of the medium in which the seismic waves propagate.

Processing of primary seismic data requires the use of high-performance technologies, that are not always available either at the datacentre's of facilities or within individual research groups. Thus there is a need for an accessible facility of "software as infrastructure" capable of "data-rich" and "cpu-rich" tasks. Such an infrastructure should provide, in a flexible and maintainable way, tools and capabilities for data and model sharing, information and knowledge extraction, data mining, processing and visualisation. Although the format of the raw data is usually represented in a well established community standard (MSEED), each datacentre can have a different data access and management system. Thus to be as flexible as possible a framework has to deal, for example, with legacy distributed databases and file system archives, heterogeneity in data and metadata storage.

In this respect ADMIRE proved itself to be very suitable with easily configurable solutions, capable of handling with large data volumes. This confirmed its main mission of delivering a professionally engineered framework, which provides a consistent and easy-to-use technology for extracting information and knowledge from data coming from multiple heterogeneous and distributed resources. High deployment flexibility and decoupling between the implementation and use of the infrastructure can been achieved thanks to the separation of concerns among the domain and the data-intensive experts, which represent the core concept on top of which the ADMIRE architecture has been built.

ADMIRE is offering a solution which is supported by a formal language (DISPEL, Data-Intensive Systems Process Engineering Language) as the end-user development interface for the underlying infrastructure. DISPEL allows composition of the data analysis steps into composite scenarios that can be considered as a set of, to some degree, independent workflows for raw data gathering and pre-processing, data aggregation, cross-correlation and stacking.

The building blocks of a DISPEL workflow are called Processing Elements (PEs), the execution of which is optimised and parallelised throughout the nodes of the ADMIRE's computational infrastructure, which relies on the OGSA-DAI data-streams processing technology. An important effort has been the dedicated integration of the domain specific API (ObsPy, http://www.obspy.org) into OGSA-DAI Activities, as the low level implementation for the seismological analysis PEs. This implementation choice was driven by a more general and important requirement of the seismology scientists who expect to develop and use their own analysis code, written with their favourite programming language and API. Such requirements lead to several considerations on the future developments which will impact on those aspects of the framework that could be improved to offer a dynamic integration of User-Defined PEs, enabling in such a way an efficient deployment and validation of user's analysis code within a deployed ADMIRE infrastructure.

Besides these interesting and applicable improvements, our experience with ADMIRE showed that the framework is already a valuable solution to keep all of the "plumbing" and the machinery hidden from the user,

while still enabling efficient, easy-to-use and transparent computational science.