

A Cloud-based Virtual Observatory for Environmental Science

Gordon S. Blair, Yehia El-khatib
School of Computing & Communications, Lancaster University, UK

Environmental scientists are increasingly being asked to answer more complex scientific questions for example related to the implications of environmental management decisions on a catchment or on the policy implications of climate change at a national or international level, a trend we refer to as 'big science' for the purposes of this paper. However, there are a number of obstacles that restrict environmental scientists from tackling such big science issues. Many of these obstacles are technical and related to different forms of fragmentation including spatial and temporal gaps in data, disrupted alignment and representation of data, unlinked models, and disjoint disciplines. Other obstacles include the difficulty of managing and processing extremely large datasets.

The EVOp project seeks to solve many of these problems by developing a virtual observatory (VO) that enables the integration of a variety of information sources (including disparate data sets, sensor data and models) at different granularities and scales. The VO will also provide interoperability with associated information services and encourage the flow from data to knowledge to policy setting in the quest for answering big science questions.

The VO requires the construction of a cyber-infrastructure that provides simplified access to data, integration with current information services, and the ability to handle large datasets. Cloud computing offers great opportunities to undertake such tasks through harnessing economies of scale offered by commodity hardware data centres. It also offers universal access to the VO that can be used by scientists, policy makers and the general public alike and, through this, achieving the desired level of integration.

Cloud computing is a distributed paradigm in which computational and storage requirements are provided in an on-demand fashion by large clusters of commodity computers. Such a pay-per-use model is generally made possible through virtualisation, i.e. using virtual machines to create custom execution environments. To the consumers of such service, they obtain customised and isolated computing resources as and when required without the need to invest in hardware that might not be fully utilised, which will depreciate in value, and will require operation, support and maintenance costs. Virtualisation also allows cloud service providers to manage large data centres at a low overall maintenance cost, and provide varying and tiered services to different customers.

Different levels of cloud services are available. Users of Software as a Service (SaaS) are able to run software through a thin client, such as a Web browser, without actually running the application on their computers. Platform as a Service (PaaS) provides a platform (i.e. a virtualised hardware setup along with an operating system) that can be used to deploy highly customised software. Finally, Infrastructure as a Service (IaaS) provides a pool of virtualised hardware resources for the user to use and manage.

There are a number of challenges involved in developing an environmental cloud, including most prominently what architecture is right to support Environmental Sciences. Hardware assets that will be required by an environmental cloud could be leased from one or more cloud service provider, such as Amazon and Google. Alternatively, dedicated data centres could be assembled to serve the needs of the VO. A hybrid approach is also possible where owned data centres would serve as primary resources while overflow requests would be redirected to third-party cloud service providers.

Second is the issue of the specific services that will be provided by the VO. Clearly, some SaaS provisioning is required to enable especially non-specialist users to access the VO. However, further services are required to enable environmental scientists to exploit the capabilities of the cloud to tackle their specific problems by mining into extremely large datasets. Adapting a distributed programming framework such as MapReduce to the needs of the environmental community is an example.

The third challenge is data discovery. Observations and models are generated by a large number of sources within and outside the scientific community. The VO will provide the potential to distinguish associations and overlaps between different datasets and models. However, there is

little incentive for many producers to share their data and models. Through the use of publishing incentives, EVOp partners will work on encouraging such producers to register their data and models. Easy-to-use desktop tools will be employed to ingest datasets and models into the cloud.

Data collected by the VO would need to be aligned to standard formats understood by different research communities, e.g. INSPIRE and data.gov. This normalisation of representation is necessary to facilitate low overhead curation, interoperability, and serendipity. The final challenge is to enable access to the data, model and visualisation tools. This will be achieved via use of a multi-faceted portal that caters for the needs of the different stakeholders, i.e. specialists, policy makers, and local communities. The use of a RESTful architecture would allow this access portal to use Web browsers that are universal client tools that are accessible to everyone and from a great deal of devices. It also allows the portal to harness common visualisation tools, such as Google Maps and Google Earth.

Cloud computing offers promising opportunities to enable the environmental community to better tackle challenges that face local communities, nation states and the international community. EVOp is a 2-year pilot project aimed to demonstrate how such technology can be used in environmental management, to help set international standards for exchanges of data and models, and to stimulate discussion within the environmental community both nationally in the UK and internationally. We are looking to collaborate with the other efforts on the regional and international scales to assist with this journey.

Acknowledgement:

The authors wish to acknowledge the Natural Environment Research Council for funding the EVOp project under grant reference NE/I002200/1. The authors also wish to thank the valuable contributions of the EVOp consortium.