Geophysical Research Abstracts
Vol. 17, EGU2015-13090, 2015
EGU General Assembly 2015
© Author(s) 2015. CC Attribution 3.0 License.

Cloud hosting of the IPython Notebook to Provide Collaborative Research Environments for Big Data Analysis

Philip Kershaw (1,6), Bryan Lawrence (1,2,5), Jose Gomez-Dans (4,6), and John Holt (3)

(1) Centre for Environmental Data Archival, STFC Rutherford Appleton Laboratory, Didcot, United Kingdom, (2) Department of Meteorology, University of Reading, Reading, United Kingdom, (3) Tessella plc., Abingdon, United Kingdom, (4) University College London, London, United Kingdom, (5) National Centre for Atmospheric Science, United Kingdom, (6) National Centre for Earth Observation, United Kingdom

We explore how the popular IPython Notebook computing system can be hosted on a cloud platform to provide a flexible virtual research hosting environment for Earth Observation data processing and analysis and how this approach can be expanded more broadly into a generic SaaS (Software as a Service) offering for the environmental sciences.

OPTIRAD (OPTImisation environment for joint retrieval of multi-sensor RADiances) is a project funded by the European Space Agency to develop a collaborative research environment for Data Assimilation of Earth Observation products for land surface applications. Data Assimilation provides a powerful means to combine multiple sources of data and derive new products for this application domain. To be most effective, it requires close collaboration between specialists in this field, land surface modellers and end users of data generated. A goal of OPTIRAD then is to develop a collaborative research environment to engender shared working.

Another significant challenge is that of data volume and complexity. Study of land surface requires high spatial and temporal resolutions, a relatively large number of variables and the application of algorithms which are computationally expensive. These problems can be addressed with the application of parallel processing techniques on specialist compute clusters. However, scientific users are often deterred by the time investment required to port their codes to these environments. Even when successfully achieved, it may be difficult to readily change or update. This runs counter to the scientific process of continuous experimentation, analysis and validation.

The IPython Notebook provides users with a web-based interface to multiple interactive shells for the Python programming language. Code, documentation and graphical content can be saved and shared making it directly applicable to OPTIRAD's requirements for a shared working environment. Given the web interface it can be readily made into a hosted service with Wakari and Microsoft Azure being notable examples. Cloud-hosting of the Notebook allows the same familiar Python interface to be retained but backed by Cloud Computing attributes of scalability, elasticity and resource pooling. This combination makes it a powerful solution to address the needs of long-tail science users of Big Data: an intuitive interactive interface with which to access powerful compute resources.

IPython Notebook can be hosted as a single user desktop environment but the recent development by the IPython community of JupyterHub enables it to be run as a multi-user hosting environment. In addition, IPython.parallel allows the exposition of parallel compute infrastructure through a Python interface. Applying these technologies in combination, a collaborative research environment has been developed for OPTIRAD on the UK JASMIN/CEMS facility's private cloud (http://jasmin.ac.uk). Based on this experience, a generic virtualised solution is under development suitable for use by the wider environmental science community - on both JASMIN and portable to third party cloud platforms.