P. Votava
NCTS# 25415-17 AGU Fall Meeting 2016

Using Docker Containers to Extend Reproducibility Architecture for the NASA Earth Exchange (NEX)

NASA Earth Exchange (NEX) is a data, supercomputing and knowledge collaboratory that houses NASA satellite, climate and ancillary data where a focused community can come together to address large-scale challenges in Earth sciences. As NEX has been growing into a petabyte-size platform for analysis, experiments and data production, it has been increasingly important to enable users to easily retrace their steps, identify what datasets were produced by which process chains, and give them ability to readily reproduce their results. This can be a tedious and difficult task even for a small project, but is almost impossible on large processing pipelines. We have developed an initial reproducibility and knowledge capture solution for the NEX, however, if users want to move the code to another system, whether it is their home institution cluster, laptop or the cloud, they have to find, build and install all the required dependencies that would run their code. This can be a very tedious and tricky process and is a big impediment to moving code to data and reproducibility outside the original system. The NEX team has tried to assist users who wanted to move their code into OpenNEX on Amazon cloud by creating custom virtual machines with all the software and dependencies installed, but this, while solving some of the issues, creates a new bottleneck that requires the NEX team to be involved with any new request, updates to virtual machines and general maintenance support. In this presentation, we will describe a solution that integrates NEX and Docker to bridge the gap in code-to-data migration. The core of the solution is saemi-automatic conversion of science codes, tools and services that are already tracked and described in the NEX provenance system, to Docker - an open-source Linux container software. Docker is available on most computer platforms, easy to install and capable of seamlessly creating and/or executing any application packaged in the appropriate format. We believe this is an important step towards seamless process deployment in heterogeneous environments that will enhance community access to NASA data and tools in a scalable way, promote software reuse, and improve reproducibility of scientific results.

Authors

- Petr Votava
  - NASA Ames Research Center
  - University Corporation at Monterey Bay
- Andrew Michaelis
  - California State University Monterey Bay
- Ryan Spaulding
  - ADNET Systems Inc. Reston
- Jeffrey C Becker
  - Computer Sciences Corporation
  - NASA Ames Research Center