ALCS Advanced Data Training Course
5/11/2019

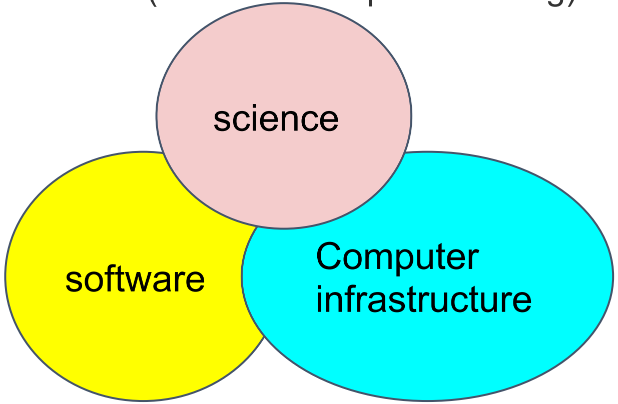
NCI AUSTRALIA

Introduction to Pangeo Environment on Raijin/Gadi

Rui Yang, Jingbo Wang, Nigel Rees

Pangeo - An open source big data geoscience platform

Pangeo is community that promotes open, reproducible, and scalable science. (NSF 2017 Sep - 2020 Aug)



This community provides

- Documentation
- Develops and maintains software
- Deploys computing infrastructure

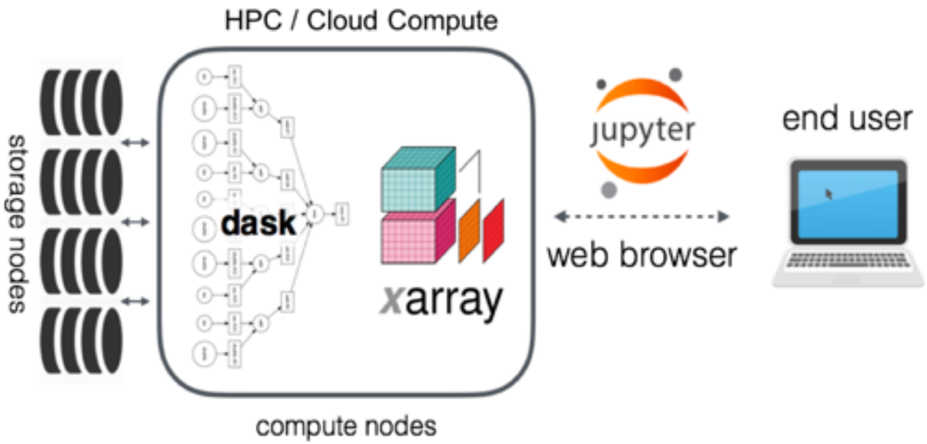
Pangeo focuses on open source tools

- Xarray, Iris, Dask, Jupyter
- many other packages

Goals:

- Foster collaboration around the open source scientific python ecosystem for ocean / atmosphere / land / climate science.
- Support the development with domain-specific geoscience packages.
- Improve scalability of these tools to handle petabyte-scale datasets on HPC and cloud platforms.

Pangeo Architecture



The key concepts and tools in the Pangeo ecosystem:

- Ability to use high-level data models (e.g. Xarray)
- Ability to leverage distributed parallel computing (e.g. Dask) on HPC systems or on cloud computing systems
- Ability to work interactively (e.g. Jupyter) or using batch processing

(https://pangeo.io/architecture.html)

NCI AUSTRALIA

Why do we need Pangeo?

“Learning more about HPC and computing with big dataset” -- from pre-training survey

A supporting community where you can find documentation, use cases for reference.

We need suitable tools for big datasets in geosciences.

Open source tools ONLY

Multiple languages support - Jupyter can be configured to run Kernels in many different languages

Pangeo is Available at NCI

Pangeo is installed on Raijin/Gadi under /apps

Users need to install extra modules if they are not included in the official pangeo environment

- Instruction could be provided.
- additional package installation requirement can be submitted via help@nci.org.au

NCI will update Pangeo versions on a quarterly basis

When to use Pangeo, when not

Pangeo is not a replacement of VDI !

Try Pangeo at Raijin/Gadi if your code

- can utilizes dask or xarray for parallel computing and data processing, and
- needs more resource such as CPU, memory and I/O throughouts.

Notice: Pangeo should be submitted to the queue system and it is recommended to request node based resources.

Don’t waste your SUs!

NCI provides VDI

- to execute lightweight, serial/parallel jupyter notebook or python script without consuming SUs.
- to develop script for Pangeo environment.
- For more details of VDI, see https://opus.nci.org.au/display/Help/VDI+User+Guide

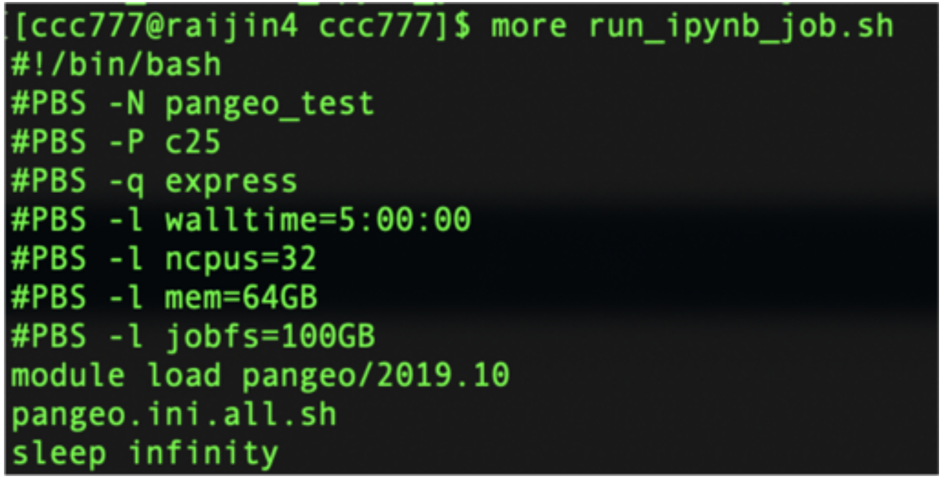
Pangeo Setup Tutorial

Tutorial: https://nci-data-training.readthedocs.io/en/latest/_notebook/general/Setup_Pangeo_environment.html

Example and relevant files are available on /g/data/c25/public, you need to copy into your own directory first.

```
>mkdir /g/data/c25/aaa777
>cp -r /g/data/c25/public/* /g/data/c25/aaa777/
>cd /g/data/c25/aaa777
```

Important Notes



```
[ccc777@raijin4 ccc777]$ more run_ipynb_job.sh
#!/bin/bash
#PBS -N pangeo_test
#PBS -P c25
#PBS -q express
#PBS -l walltime=5:00:00
#PBS -l ncpus=32
#PBS -l mem=64GB
#PBS -l jobfs=100GB
module load pangeo/2019.10
pangeo.ini.all.sh
sleep infinity
```

- Make sure your project has enough ksu
- In the queue, normal is recommended if not urgent
- Walltime provide the limit of the job run, make sure you have enough walltime to run an computationally expensive job
- Watch for the new version of Pangeo in three months time
- ...

Important Notes

Add those lines into your notebook or python script!!

```
# start the dask client
 from dask.distributed import Client,LocalCluster
  client =  Client(scheduler_file='scheduler.json')

… your work utilizing xarray&dask

# stop the pbs job.
! pangeo.end.sh
```

What if I need additional libraries?

Install pangeo compatible modules at Raijin/Gadi

Step 1. Load pangeo module

```
bash-4.1$ module load pangeo/2019.10
bash-4.1$ source ${PANGEO_ROOT}/etc/profile.d/conda.sh
bash-4.1$ conda activate pangeo
```

Install pangeo compatible modules at Raijin/Gadi

Module deepgraph is unavailable from the default pangeo

```
(pangeo) bash-4.1$ python
Python 3.7.3 | packaged by conda-forge | (default, Jul  1 2019, 21:52:21)
[GCC 7.3.0] :: Anaconda, Inc. on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> import deepgraph
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
ModuleNotFoundError: No module named 'deepgraph'
>>> exit()
```

Install deepgraph to user's own directory

```
(pangeo) bash-4.1$ pip install --install-option="--prefix=YOUR_OWN_DIRECTORY" deepgraph
/apps/pangeo/2019.10/envs/pangeo/lib/python3.7/site-packages/pip/_internal/commands/install.py:243: UserWarning: Disabling all use of wheels due to the use of --build-options / --global-options / --install-options.
  cmdoptions.check_install_build_global(options)
Collecting deepgraph
  Downloading https://files.pythonhosted.org/packages/fc/3e/4a34a5316a5f886b8d7a6787c24852d9e5a5ef00b4ec6af0736f681a3a58/DeepGraph-0.2.2.tar.gz (160kB)
                |████████████████████████████████| 163kB 4.7MB/s
Requirement already satisfied: numpy>=1.6 in /apps/pangeo/2019.10/envs/pangeo/lib/python3.7/site-packages (from deepgraph) (1.17.2)
Requirement already satisfied: pandas>=0.17.0 in /apps/pangeo/2019.10/envs/pangeo/lib/python3.7/site-packages (from deepgraph) (0.25.1)
Requirement already satisfied: python-dateutil>=2.6.1 in /apps/pangeo/2019.10/envs/pangeo/lib/python3.7/site-packages (from pandas>=0.17.0->deepgraph) (2.8.0)
Requirement already satisfied: pytz>=2017.2 in /apps/pangeo/2019.10/envs/pangeo/lib/python3.7/site-packages (from pandas>=0.17.0->deepgraph) (2019.2)
Requirement already satisfied: six>=1.5 in /apps/pangeo/2019.10/envs/pangeo/lib/python3.7/site-packages (from python-dateutil>=2.6.1->pandas>=0.17.0->deepgraph) (1.12.0)
Skipping bdist_wheel for deepgraph, due to binaries being disabled for it.
Installing collected packages: deepgraph
  Running setup.py install for deepgraph ... done
Successfully installed deepgraph
```

Setup environment and validate

Setup environment of deepgraph

```
(pangeo) bash-4.1$ export PYTHONPATH=$PYTHONPATH:YOUR_OWN_DIRECTORY/lib/python3.7/site-packages
```

Validate deepgraph installation

```
(pangeo) bash-4.1$ python
Python 3.7.3 | packaged by conda-forge | (default, Jul  1 2019, 21:52:21)
[GCC 7.3.0] :: Anaconda, Inc. on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> import deepgraph
>>> exit()
```