JupyterHubs and containerization to streamline and accelerate science innovation
May 17, 2024 EDMW 2024

NOAA Fisheries Open Science

Elizabeth Eli Holmes, Ph.D
NMFS Open Science, Lead

Since 2022, working with NASA DAAC (data centers) colleagues on open infrastructure for geospatial analysis for cloud-native data

NOAA Fisheries Open Science JupyterHub

NOAA Fisheries Open Science

Login to the JupyterHub

Sign in with GitHub

Funded by the NOAA and Microsoft Cooperative Research and Development Agreement (CRADA) to help advance NOAA's mission to create a Climate-Ready Nation. This JupyterHub is used for internal NOAA training and workshops in cloud computing and scientific programming to support this mission. These trainings focus on analysis with geospatial earth climate data and output from climate models, processing of acoustics data from passive and active surveys, and bioinformatics toolsets for analysis of fisheries genetic data.

NOAA FISHERIES

Containerization

Install MacOS
Install R
Install RStudio
Install packages
Install Python
Install VSCode

Unbuntu
R & RStudio
packages
Python & VSCode
Packages
Extra binaries

One person or a science community creates the container image

jupyterhub

No one installs anything. You run the container.

JupyterHubs: a multi-user platform for running containers/images

OS
pangeo +
machine-learning
tools + one of our
python modules

OS
rocker/geospatial
with R 4.2 + suite
of packages

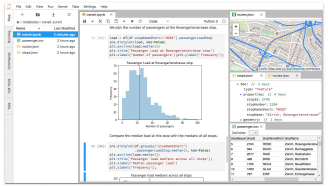
OS
Python + R +
C++
Atlantis model

microbiome
bioinformatics

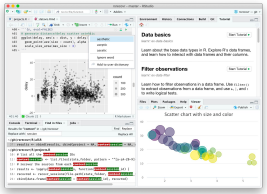
acoustics
pipelines

Container is all our packages and the os environment set up

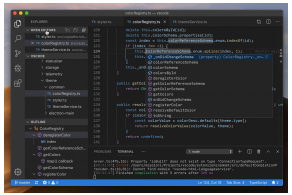
jupyterhub



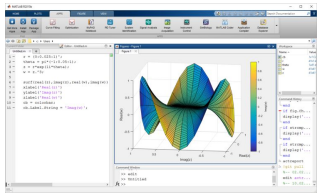
jupyterlab



rstudio



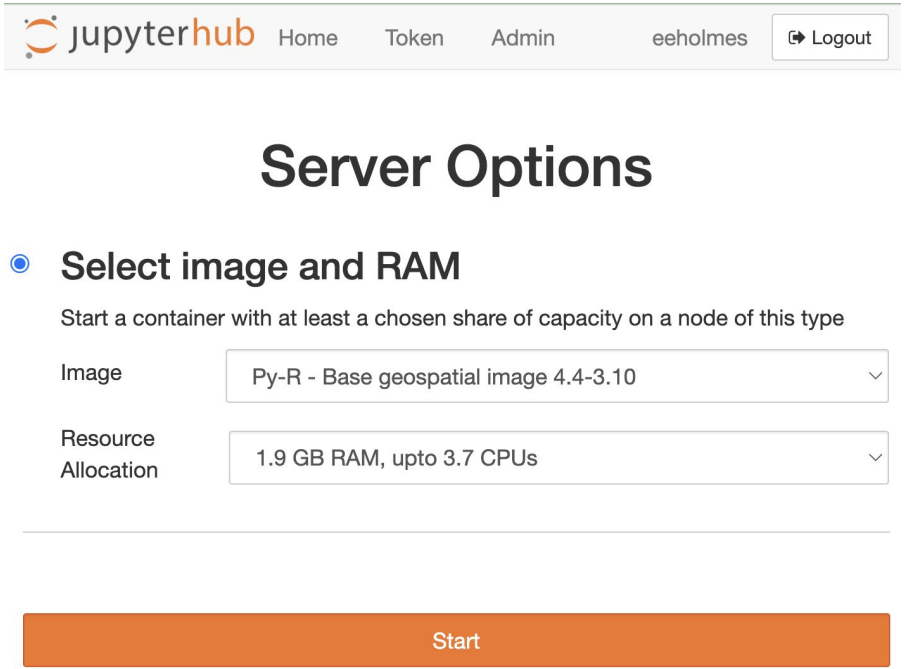
vscode

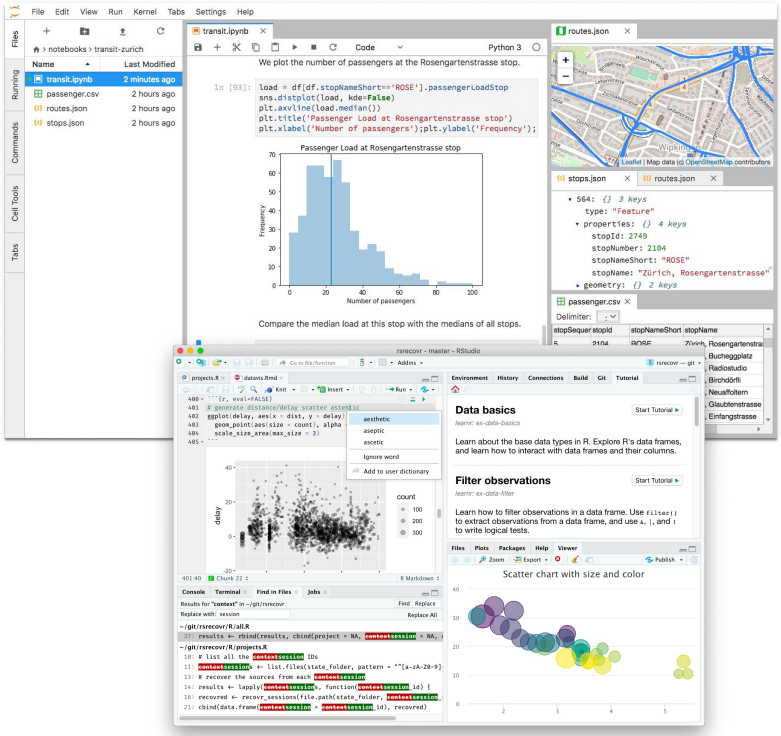


matlab

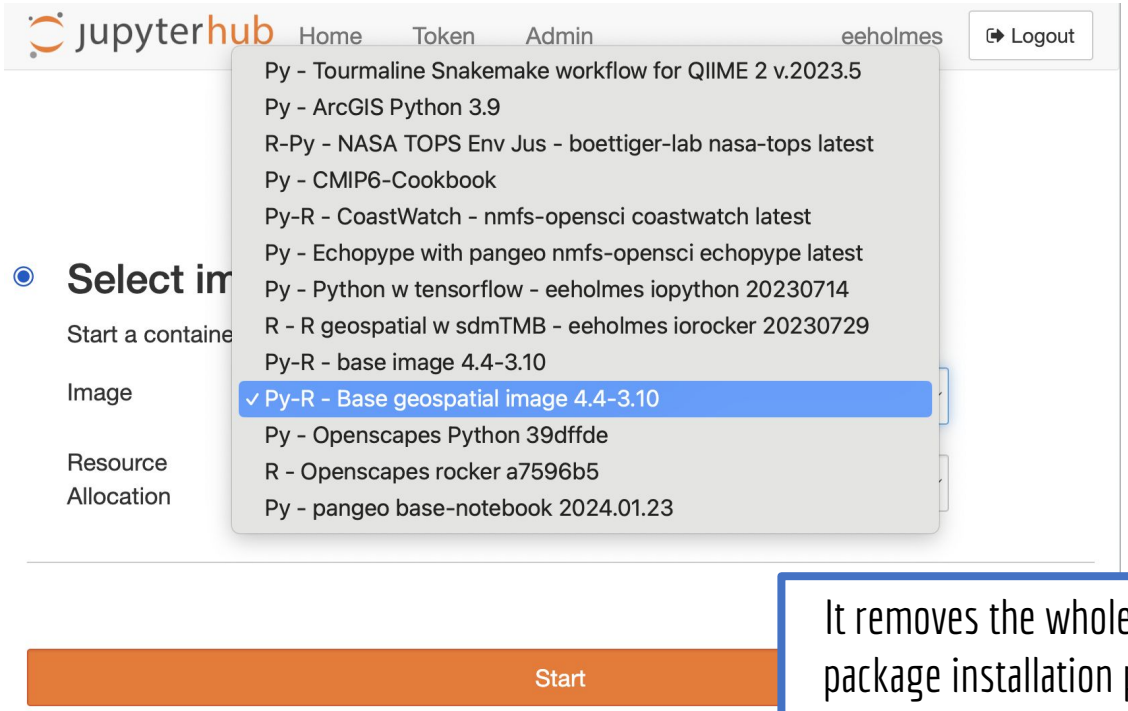
Why are JupyterHubs so popular?

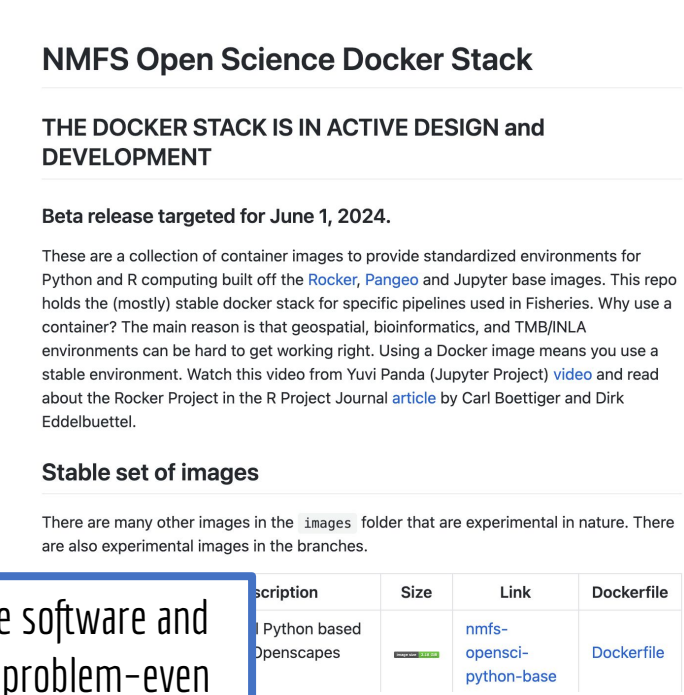
User experience is familiar and spin-up easy





Share complex compute environments easily





NMFS Open Science Docker Stack

THE DOCKER STACK IS IN ACTIVE DESIGN and DEVELOPMENT

Beta release targeted for June 1, 2024.

These are a collection of container images to provide standardized environments for Python and R computing built off the Rocker, Pangeo and Jupyter base images. This repo holds the (mostly) stable docker stack for specific pipelines used in Fisheries. Why use a container? The main reason is that geospatial, bioinformatics, and TMB/INLA environments can be hard to get working right. Using a Docker image means you use a stable environment. Watch this video from Yuvi Panda (Jupyter Project) video and read about the Rocker Project in the R Project Journal article by Carl Boettiger and Dirk Eddelbuettel.

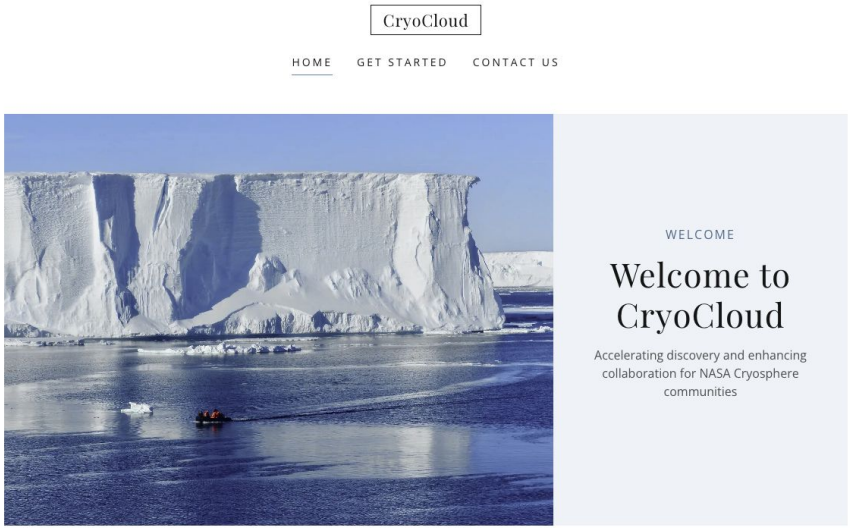
Stable set of images

There are many other images in the `images` folder that are experimental in nature. There are also experimental images in the branches.

scription	Size	Link	Dockerfile
l Python based Openscapes		nmfs-opensci-python-base	Dockerfile

It removes the whole software and package installation problem–even for complex environments

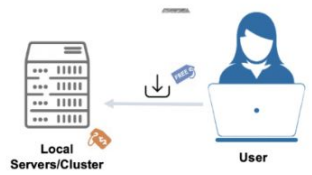
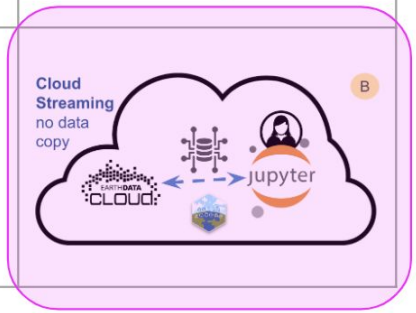
JupyterHubs can accelerate collaboration



CryoCloud example (Tasha Snow, PI). ICESat-2 community

James Colliander (presenting for PI Tasha Snow at JupyterCon 2023): Accelerating Discovery for NASA Cryosphere Communities with JupyterHub

JupyterHubs allow you to get “next” to the data

		Compute	
		Not Cloud	Cloud
Data storage	Not Cloud		
	Cloud		

Curtesy of Alexis Hunzinger (NASA Openscapes)

*if all data were cloud-optimized it might not matter, but it’s not.

Data IO is often the bottleneck. Being next to the data buckets means you can “attach” a huge cloud drive to your virtual computer*.

Cloud-native provides new options for massively parallel computing.

JupyterHubs use Open Infrastructure

Infrastructure (JupyterHub) is cloud-provider agnostic and can be run on ‘bare-metal’ or even a laptop.

Base “Images” are infrastructure agnostic. Can be dropped into:
Jupyter server,
docker/podman (offline),
Binder, GitHub Codespaces,
GitLab

How does one run a JupyterHub?

Common ways people run JupyterHubs

Self-hosted hubs on ‘bare-metal’ servers or own cloud account

managed hubs

Institution managed hubs

Easy creation: A JupyterHub on Kubernetes on GCP in 5 minutes

```
# Open a Cloud Shell and run these lines of code
gcloud container clusters create \
  --machine-type n1-standard-2 \
  --num-nodes 2 \
  --zone us-west1-a \
  --cluster-version latest \
  jhub

kubectl create clusterrolebinding cluster-admin-binding \
  --clusterrole=cluster-admin \
  --user=yourname@gmail.com

gcloud beta container node-pools create user-pool \
  --machine-type n1-standard-2 \
  --num-nodes 0 \
  --enable-autoscaling \
  --min-nodes 0 \
  --max-nodes 3 \
  --node-labels hub.jupyter.org/node-purpose=user \
  --node-taints hub.jupyter.org_dedicated=user:NoSchedule \
  --zone us-west1-a \
  --preemptible \
  --cluster jhub
```

```
curl
https://raw.githubusercontent.com/helm/helm/HEAD/scri
pts/get-helm-3

helm repo add jupyterhub https://hub.jupyter.org/helm-chart/
helm repo update

helm upgrade --cleanup-on-fail \
  --install jhub1 jupyterhub/jupyterhub \
  --namespace jhubk8 \
  --create-namespace \
  --version=3.3.4 \
  --values config.yaml

# DONE!!!
```

Cost for running your own JupyterHub. My experience.

In the cloud for 3 people full-time
$500 / month

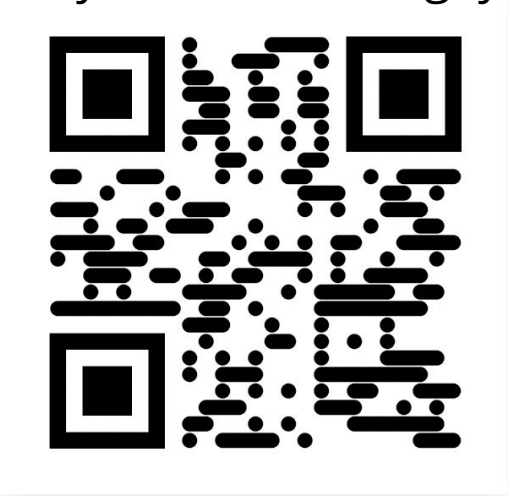
But there is the time to set-up. About a week of futzing if you have never done it. 2-4 hrs of futzing otherwise.
AWS, Google, Azure, any other cloud provider

Large community JupyterHub focused on geospatial analysis

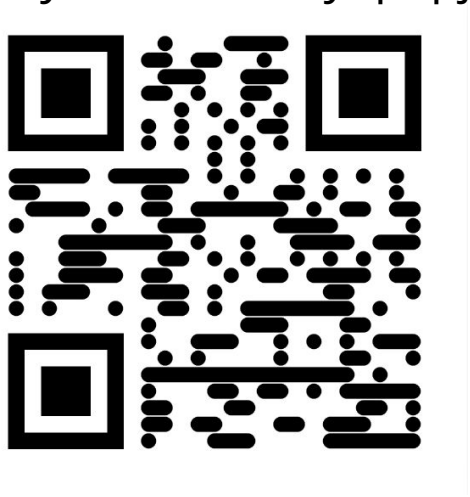
- $2500/mo admin and support (outside)
- $1000-1500/mo compute + storage

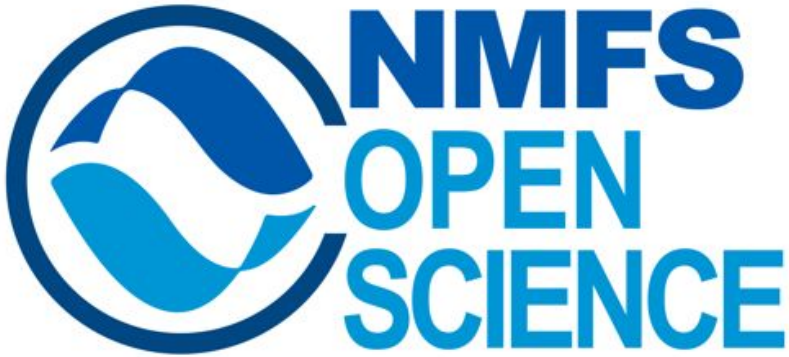
See a JupyterHub Demo

Basic
https://youtu.be/S-OXTg9yadQ



Bring your own image
https://youtu.be/WvyepapjTEE





https://nmfs-opensci.github.io/

Look for link to our internal site!