

HelioCloud

Brian Thomas (NASA HDRL), Sandy Antunes (APL)

IHDEA, October 2023

Heliophysics data landscape:

- **Big Datasets** enable new science
 - ... but too big to load on your laptop
 - petabytes, millions of files, decade-long timelines
 - requires significant compute capability to use effectively
- Community has **workflow they like**:
 - old way: *search, fetch, download, analyze on laptop*
 - cloud approach is analyze on remote machine, 'science in the browser'
 - data in cloud should use same search portals, file types, code libraries

HDRL? Who?

HP Data and Model Consortium

Brian Thomas (Acting PS)

Overall management of the HDRL.

Registries and DOIs for all digital resources; SPASE Data Model.

Heliophysics Data Portal (HDP; including solar)

Python and other software integration (PyHC).

Analysis and visualization services ((Py)SPEDAS, Autoplots).

Data upgrades and services.

HelioCloud initiative with data and software from all groups.

Space Physics Data Facility

Robert Candey (PS), Lan Jian (DPS)

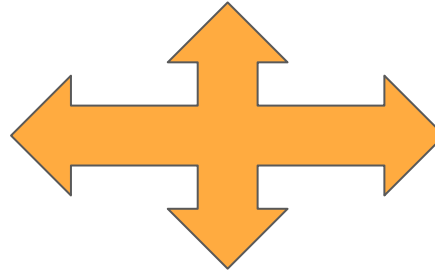
Non-solar Data Final Active Archive for NASA (and other) missions.

CDAWeb data browsing and access; Web Service access.

OMNIWeb data production and serving.

SSCWeb and 4-D spacecraft orbit facility.

Common Data Format.



Solar Data Analysis Center

Jack Ireland (PS)

Solar Data Final Active Archive for Solar Dynamics Observatory

and other NASA missions.

Virtual Solar Observatory data access.

Heliviewer. SunPy. SolarSoft.

High Performance Computing for NASA HP.

Collaborators

Community Coordinated Modeling Center

Center for HelioAnalytics

Data—model comparisons. Registry of models and output.

“Kamodo” enabled visualization. AI/ML cloud computing.

Analytics tools and methods.

HelioCloud - AWS Cloud Compute + Big Data + GUIs

Several capabilities available to researchers

- **Easy Collaboration via Notebooks** ("science analysis via GUI")
- **Direct access to AWS's arcane compute, costing**
- **Big cheap storage**

Example: 2TB of SDO EUV data

Here we combine a few HelioCloud demos with our big data fileRegistry to tackle 1 year of SDO data. The data is in AWS S3 and we do not copy the files over, but instead have CPUs at AWS access it directly

1 year of SDO 94A EUV images from AIA is 129,758 files, each 14MB, totalling 1.8 TB.

This code calculates a simple irradiance change over time

If done serially on your laptop, it would take 27 hours

HelioCloud takes 25 minutes (1467 sec) to analyze through 1 year of SDO data

More fun stats: that's 88 files/second, also 1 GB/sec to do the full analysis (which is 2x the read speed of a SATA SSD). It is 2x faster than it would take to just copy the files off a local disk and 8x faster than copying via gigabit internet.

For a given S3 location, list of all available datasets

```
fr=scregistry.FileRegistry("s3://gov-nasa-hdrl-data1/")
# Now we get metadata associated with AIA 94A
frID = "aia_0094"
myjson = fr.get_entry(frID)
start, stop = myjson['start'], myjson['stop']
```

Now let us get the entire list of SDO files for that AIA ID

```
file_registry1 = fr.request_file_registry(frID, start_date=start, stop_date=stop, overwrite=False)
# And convert that richer data to a list of files to process
filelist = file_registry1['datakey'].to_list()
```

```
file_registry1
```

	start	datakey	filesize
0	2020-02-17 00:00:00	s3://gov-nasa-hdrl-data1/sdo/aia/20200217/0094...	13910400
1	2020-02-17 00:04:00	s3://gov-nasa-hdrl-data1/sdo/aia/20200217/0094...	13910400

```
def DO_SCIENCE(mydata):
    # you can put better science here
    iirad = mydata.mean()
    return iirad
```

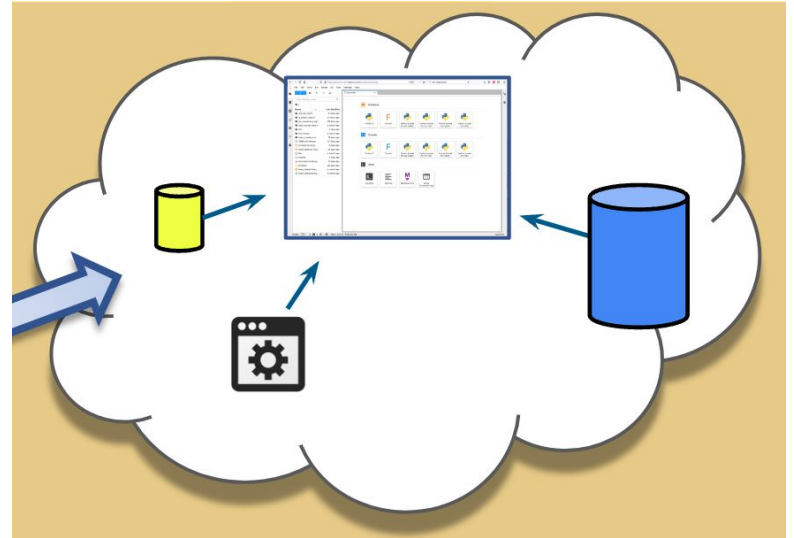
Vision / Goals

- Create *reusable software stack* for Heliophysics Research
 - Useful for *ground-based and space-based research*
 - Useful for *high-school to professional research* use
- Create extensible, reusable research environment
 - Used locally 'on laptop' or used in cloud computing, with ability to 'burst'
- Create easily deployed cloud-based platform
 - Has additional services which help with Team collaboration, cloud-based research
 - AWS first, but could potentially be adapted to other cloud infrastructure

Inspiration : PyHC (<https://heliopython.org>), Pangeo (<https://pangeo.io>), IRAF (<https://iraf-community.github.io/>)

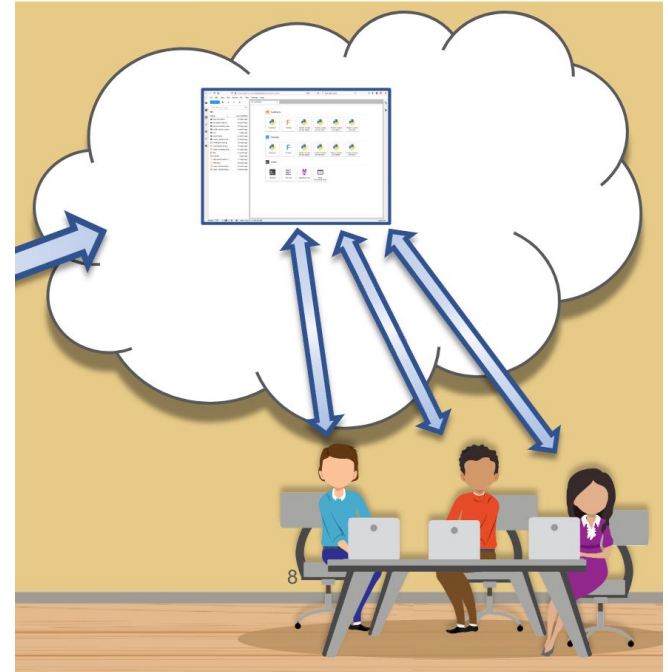
Use Case: Big Data Science

- ***Easy to use platform*** (based on Pangeo/JupyterHub)
- ***Compute “Bursts” in the cloud*** to make more processing available
- ***Multi-PB science data volumes*** resident in the cloud (AWS)



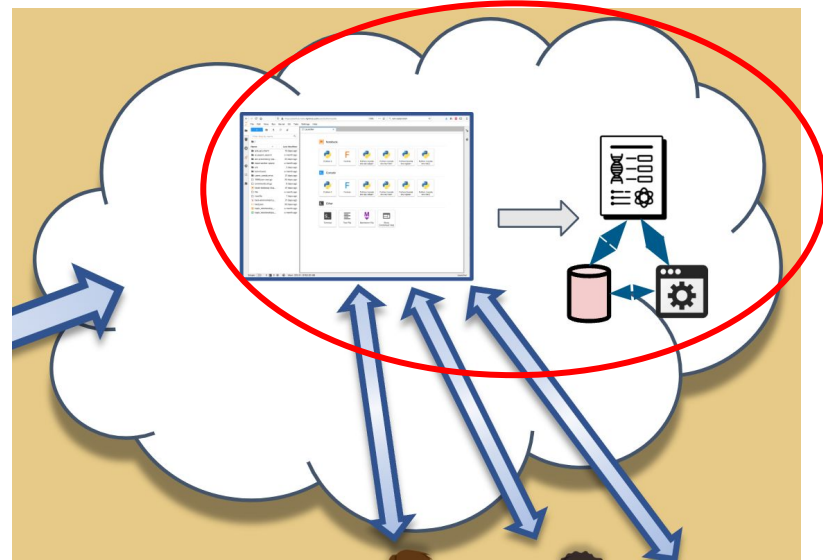
Use Case : Collaboration

- **Everyone has access** to the same environment (no 'it doesn't work for me problem')
- **Pre-installed heliophysics science software** (ex. PyHC)
- **Easily extensible software** environment by scientists
- **Easier and faster onboarding** for some institutions (e.g. NASA)



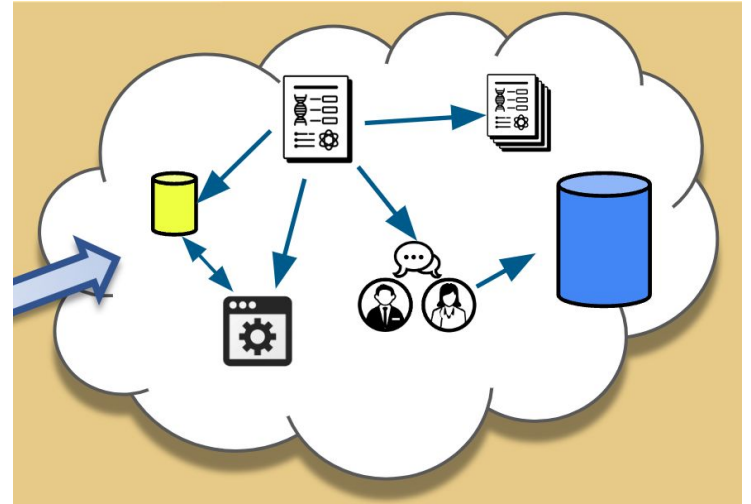
Use Case: Open Science Publishing

- ***Tools to provide metadata*** required for interlinking and discovery (SPASE model)
- ***Tools for publishing team data, notebooks and software*** (containers) to cloud for others to discover and use.
- ***Tools, Tutorials / Instructions for publishing with major journals.***



Use Case : Discovery of Interlinked Research Artifacts

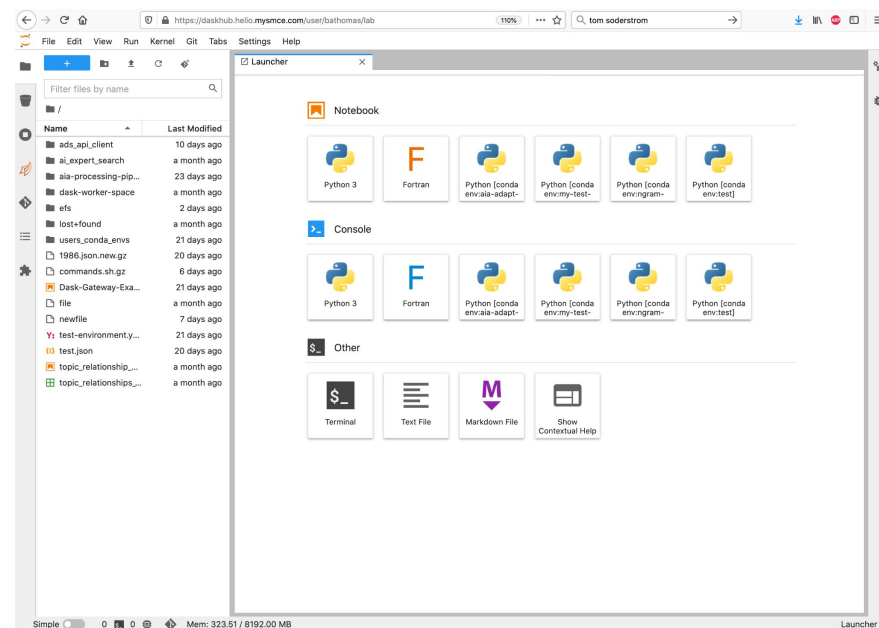
- **Leveraging DOIs** for datasets, code, people, and even more 'exotic' things.
- **Interlinking with literature** (Working with ADS; Astrophysical Data Service)
- **Enabling better discovery** by developing dictionaries, services and tools



HelioCloud platform: Easy Cloud Onramp for Science

Allows for both Jupyter notebook and terminal-based work

- **Scientific computing focus**
 - Python, IDL, Fortran, C/C++ & libraries
 - AI/ML & GPU tooling
- **Bursting**
 - Create programs which stand up powerful virtual machines (GPU, loads of RAM, etc)
- **Default software install (“Just works”)**
 - Conda (Python) environment for HelioPhysics / ML
 - Other research packages
- **Tutorials**
 - How to do a number of basic things
- **Access to large heliophysics datasets**
 - ~8 Pb by 2025
 - Mission, 3rd Party/User Published, “ML Ready”, EPO/Hackathon/Citizen Sci. Datasets
- **Has 125+ science users so far**
 - also 80+ simultaneous student users at PyHC Summer School



User Dashboard

User manages access to VMs in cloud

- Start/Stop/Terminate
- Launch VMs
- Track User Costs
- Generate SSH Keys

New to the portal? Check out our [quickstart](#) guide to get started!

Accumulated Spend

\$144.45

Spend Over Last 7 Days

\$2.54

Running Instances

No instances currently running.

Launch an Instance

Previously Stopped Instances

Instance Name: CDAWebfetch

OS Platform: Amazon Linux

Image: amzn2-ami-kernel-5.10-hvm-2.0.20220606.1-x86_64-gp2

Type: t2.medium

Actions ▾

HelioCloud Data Cache

- **High Level Data**

- SPDF : ALL CDAWeb (level 2+) : ~250 TB
- SDAC: ALL non-SDO (level 1+) : ~350 TB
- SDO : Selected AIA/HMI (level “1.5”): ~ 1 PB
- Contributed : AI/ML, other data : as space allows

- **Locations**

- S3 buckets (eventually several): “**s3://gov-nasa-hdrl-data1/**”
- Open and free to all

- **Status**

- ~350 TB SDO data are public
- SPDF data to began moving to public this semester, Finish Oct 1st 2023
- SDAC data to began moving to public this semester, Finish Dec 31st 2023
- Contributed dataset: SDMOML (FDL data product) now public

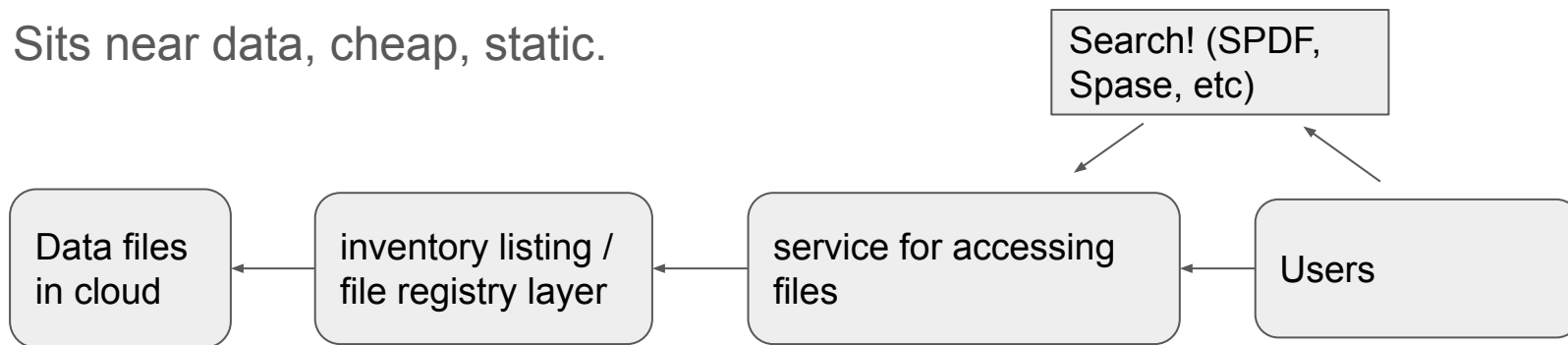
Shared Cloud Registry: a cloud-neutral alternative to 'ls -r'

'scregistry' takes a HAPI-like minimalist approach

CSV inventory listings, one per year per dataset, with time + file location + file size

Per dataset, not per site or bucket or owner.

Sits near data, cheap, static.



*.fits, *.cdf, *.h5, etc

e.g. scregistry, DB

e.g. HAPI, VirES, etc

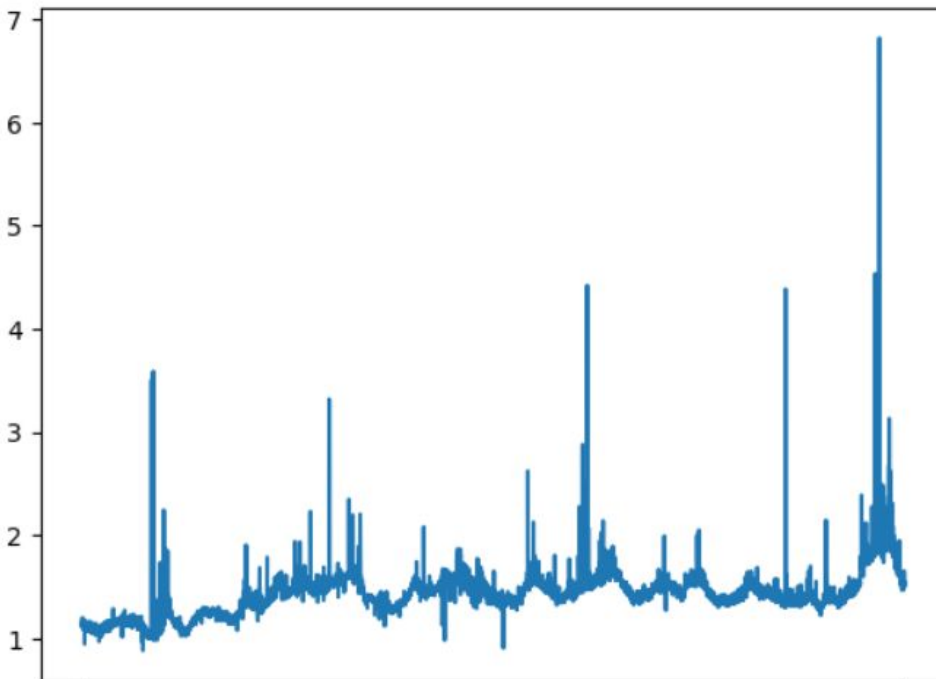
e.g. scientists, code, slurm, etc

25 minutes
later

...

Results!

```
[38]: from matplotlib import pyplot as plt
      %matplotlib inline
      # Have Matplotlib create vector (svg) instead of raster (png) images
      %%config InlineBackend.figure_formats = ['svg']
      #plt.figure()
      plt.plot(*zip(*plotme))
      plt.show()
```



Also scaling...
 10K files = 5 min
 130K files/1.8 TB in 25
 min (non-optimized)

HelioCloud - Next Year

- **Uploading ~2 PB of Data (NASA Instance)**

- Mission (SPDF IV2, SDAC selected)
- Contributed research data (ex. 'ML ready')
- Searchable

- **Facilitate HelioCloud to HelioCloud**

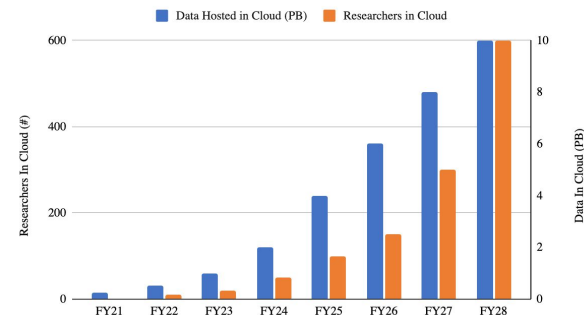
Teams may use compute at different institutions but share resources across institutional boundaries easily

- File Registry (public)
- Searchable Science Database (public)
- Notebook sharing

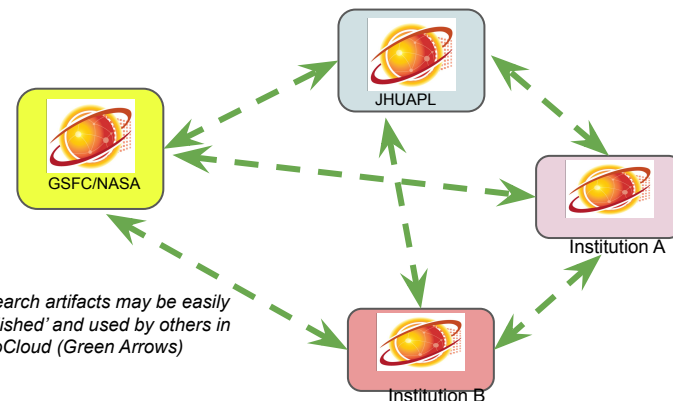
- **Even more capabilities**

- Machine Learning / GPU accelerated Dask jobs
- Private team resources (S3 buckets)
- "Container Registry" - publish and share docker containers w/ public
- "HPC Terminal" - create SLURM jobs and run on a 'virtual cluster'
- CI/CD GitLab runner (??) - incorporate best practices into your software development

HelioCloud Capability Growth



Expected Growth in HelioCloud users at GSFC / NASA. As other institutions participate, this will greatly increase beyond this prediction.



Research artifacts may be easily 'published' and used by others in HelioCloud (Green Arrows)

International Collaboration (from IHDEA 2022)

- “Organizational” : *define best practices, guidelines for cloud-based research*
 - Standard formats, services, access, registries
 - Open Data, Open Science, Persistent Identifiers for user research data, etc
 - Host international meetings to share information, build collaboration
- Coordinate, co-develop ‘PanHelio’ : *shared software environment / container*
 - IDHEA ‘sub working group’? - to oversee content in software environment
- Coordinate, collaborate on ‘Cloud’ : *cloud-based research tools & services*
 - Identify areas of ‘easy integration’ and / or critical integration (ex. data sharing, open science publication tools)

HelioCloud Team and ESAC are discussing sharing beta access, plan to tag up next in October / November

2023 HelioCloud release:

heliocloud@groups.io

Github release at AGU Dec 2023.

Extra slides

Overview

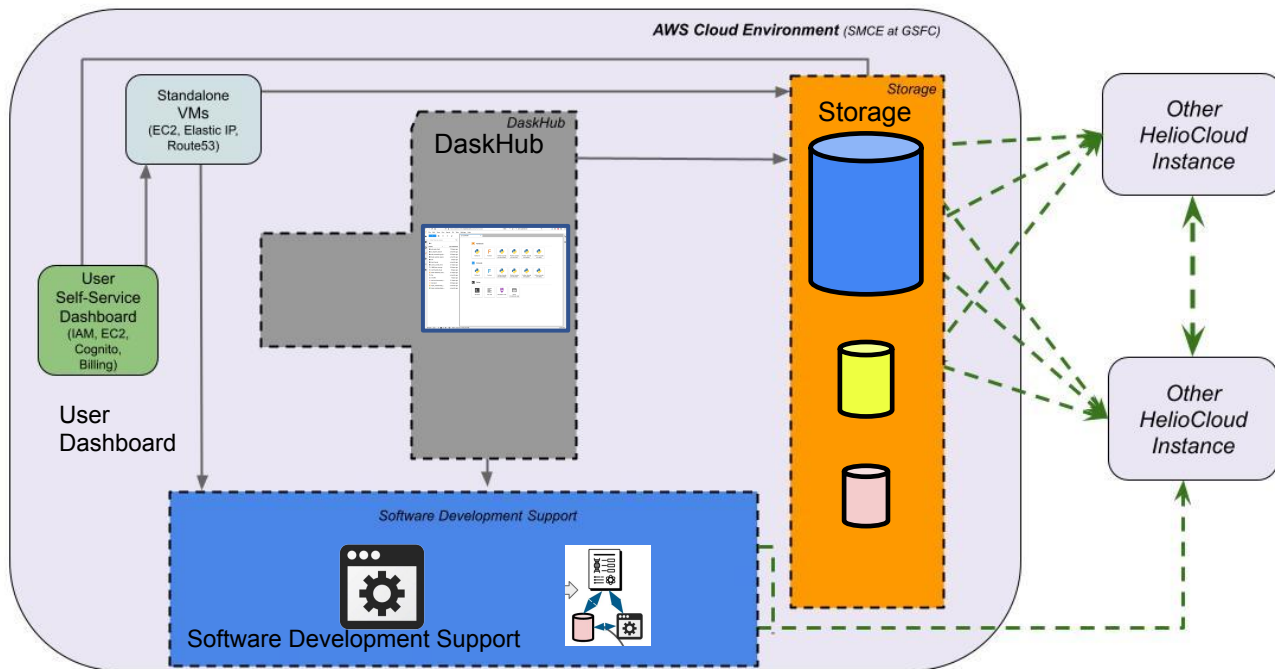
Heliophysics (HDRL) in SMCE:

- Heliophysics infrastructure
 - HelioCloud - science platform for heliophysics research
 - HDRL Infrastructure
 - Development of Helioviewer, HDP, etc.
 - Helionauts
 - Registration
 - Other HSD Infrastructure (HDRL supports)
 - Host sites for LWSTM/LWS, CfHA
- 'HelioCloud' - science platform
 - 115+ users registered, ~50 active
 - Daskhub / User Portal

HelioCloud Cloud Platform Architecture (AWS-based)

Shared Open Architecture

- Instances (purple shapes) deployed at various institutions
- Optional components
- Data, Service resources easily found, used across instances
- Costs handled by deploying institution
- Open source, community based development



High-level Architecture Diagram. HelioCloud consists of multiple optional components. Hosting entity chooses which components to deploy. Storage and Software Development Support components naturally integrate with same components deployed at other HelioCloud instances.



Researcher Use Cases for a HelioCloud environment

- **“Big Data”**: handling egress, open data

Access to significant data and compute resources

- **Open Science: enabling scientists (not dev-focus)**

Common tools to facilitate reproducibility / transparency and publication

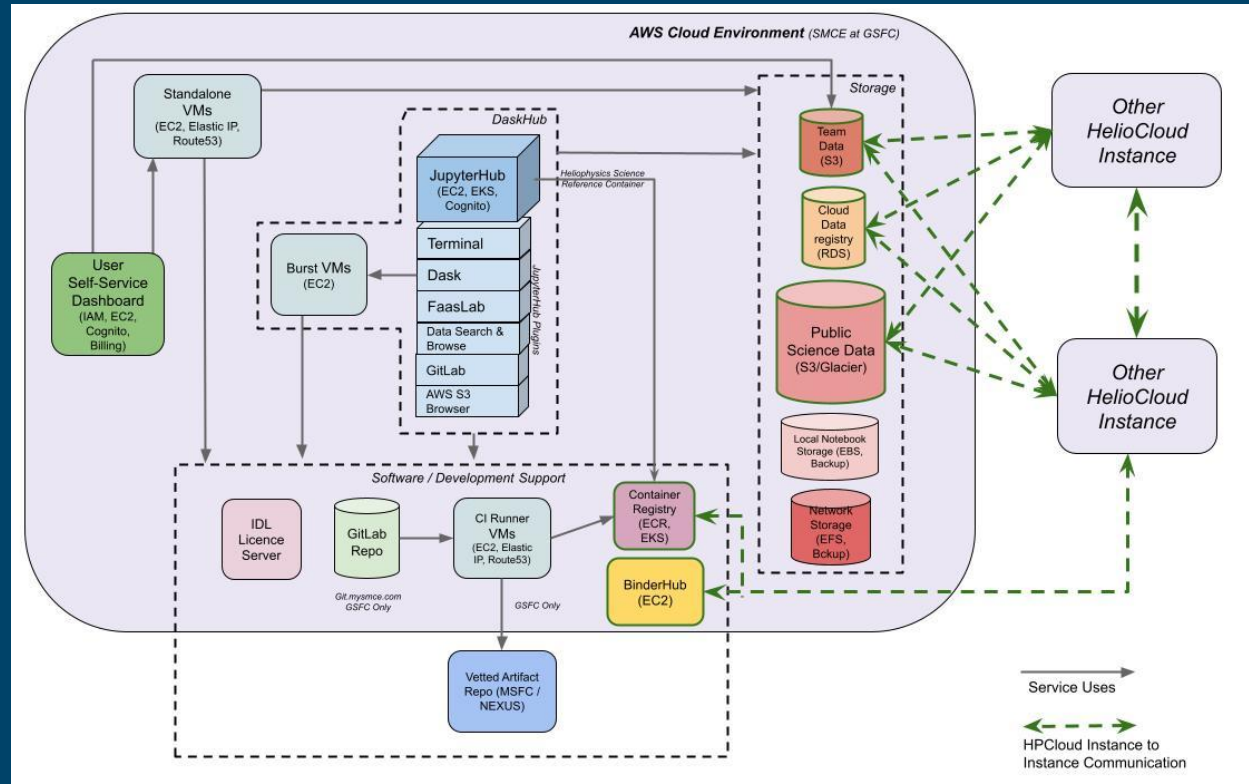
- **Collaboration: open source cloud modules, cross-cloud compute**

Shared environment easily accessed via browser, containers, etc. Facilitate science teams.

HelioCloud Architecture : More Details

Shared Open Architecture

- Instances (puce boxes) deployed at various institutions
- Optional components
- Data, Service resources easily found, used across instances
- Costs handled by deploying institution
- Open source, community based development



User Dashboard

User manages access to VMs in cloud

- Start/Stop/Terminate
- Launch VMs
- Track Costs
- Generate SSH Keys
- *Storage Provisioning (S3)*

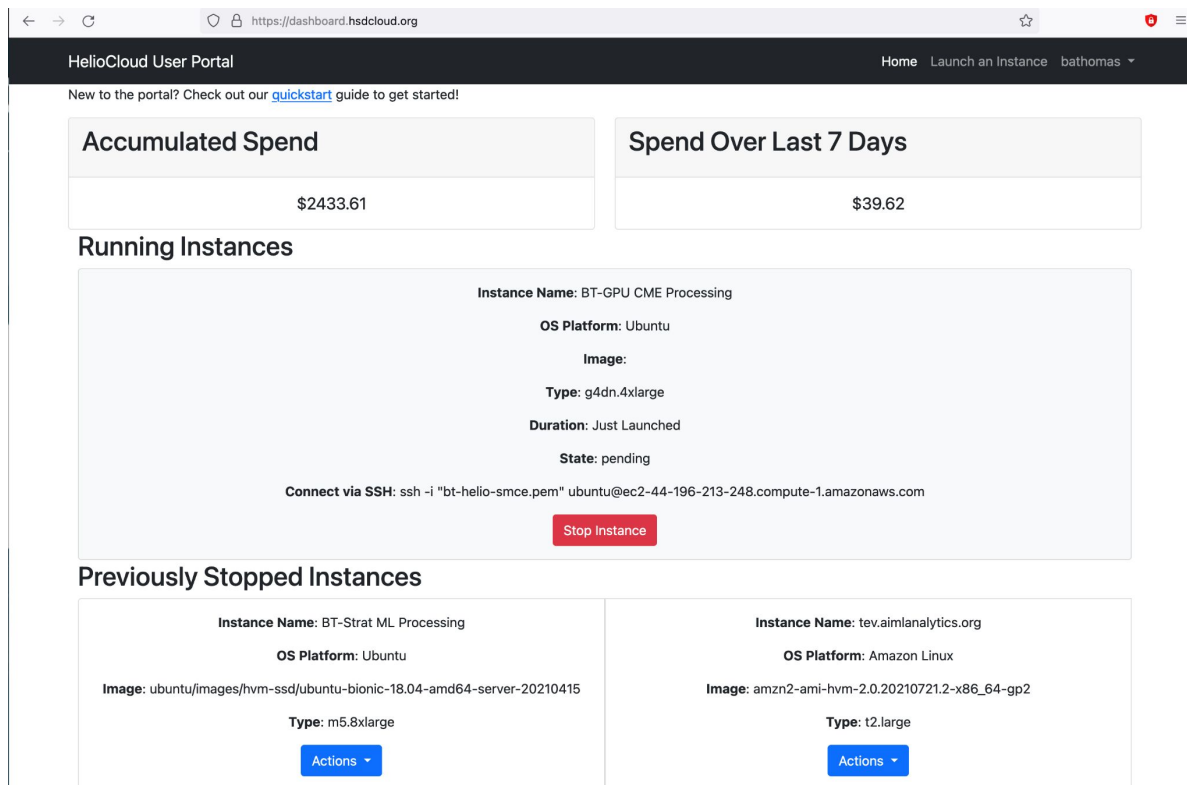
The screenshot displays the HelioCloud User Portal interface. At the top, there is a navigation bar with the title 'HelioCloud User Portal' and links for 'Home', 'Launch an Instance', and a user profile 'bathomas'. Below the navigation bar, a message reads: 'New to the portal? Check out our [quickstart](#) guide to get started!'. The dashboard is divided into several sections:

- Accumulated Spend:** A card showing a total spend of \$2433.61.
- Spend Over Last 7 Days:** A card showing a total spend of \$39.62.
- Running Instances:** A section containing one instance card for 'BT-GPU CME Processing'. The instance details are: OS Platform: Ubuntu, Image: (blank), Type: g4dn.4xlarge, Duration: Just Launched, State: pending. A red 'Stop Instance' button is visible at the bottom of the card. Below the instance details, there is a SSH connection command: `ssh -i "bt-helio-smce.pem" ubuntu@ec2-44-196-213-248.compute-1.amazonaws.com`.
- Previously Stopped Instances:** A section containing two instance cards. The first is 'BT-Strat ML Processing' with OS Platform: Ubuntu, Image: ubuntu/images/hvm-ssd/ubuntu-bionic-18.04-amd64-server-20210415, and Type: m5.8xlarge. The second is 'tev.aimlanalytics.org' with OS Platform: Amazon Linux, Image: amzn2-ami-hvm-2.0.20210721.2-x86_64-gp2, and Type: t2.large. Both cards have a blue 'Actions' button at the bottom.

User Portal : Self Service VMs in Cloud (open source AWS CDK module)

User manages access to VMs in cloud

- Start/Stop/Terminate
- Launch VMs
- Track User Costs
- Generate SSH Keys



The screenshot shows the HelioCloud User Portal interface. At the top, there is a navigation bar with 'Home', 'Launch an Instance', and 'bathomas'. Below the navigation bar, a message reads: 'New to the portal? Check out our [quickstart](#) guide to get started!'. The main content area is divided into several sections:

- Accumulated Spend:** \$2433.61
- Spend Over Last 7 Days:** \$39.62
- Running Instances:** A card for 'Instance Name: BT-GPU CME Processing' with details: OS Platform: Ubuntu, Image: (blank), Type: g4dn.4xlarge, Duration: Just Launched, State: pending. It includes an SSH command: `ssh -i "bt-helio-smce.pem" ubuntu@ec2-44-196-213-248.compute-1.amazonaws.com` and a 'Stop Instance' button.
- Previously Stopped Instances:** Two cards are shown:
 - Instance Name: BT-Strat ML Processing, OS Platform: Ubuntu, Image: ubuntu/images/hvm-ssd/ubuntu-bionic-18.04-amd64-server-20210415, Type: m5.8xlarge, with an 'Actions' button.
 - Instance Name: tev.aimlanalytics.org, OS Platform: Amazon Linux, Image: amzn2-ami-hvm-2.0.20210721.2-x86_64-gp2, Type: t2.large, with an 'Actions' button.

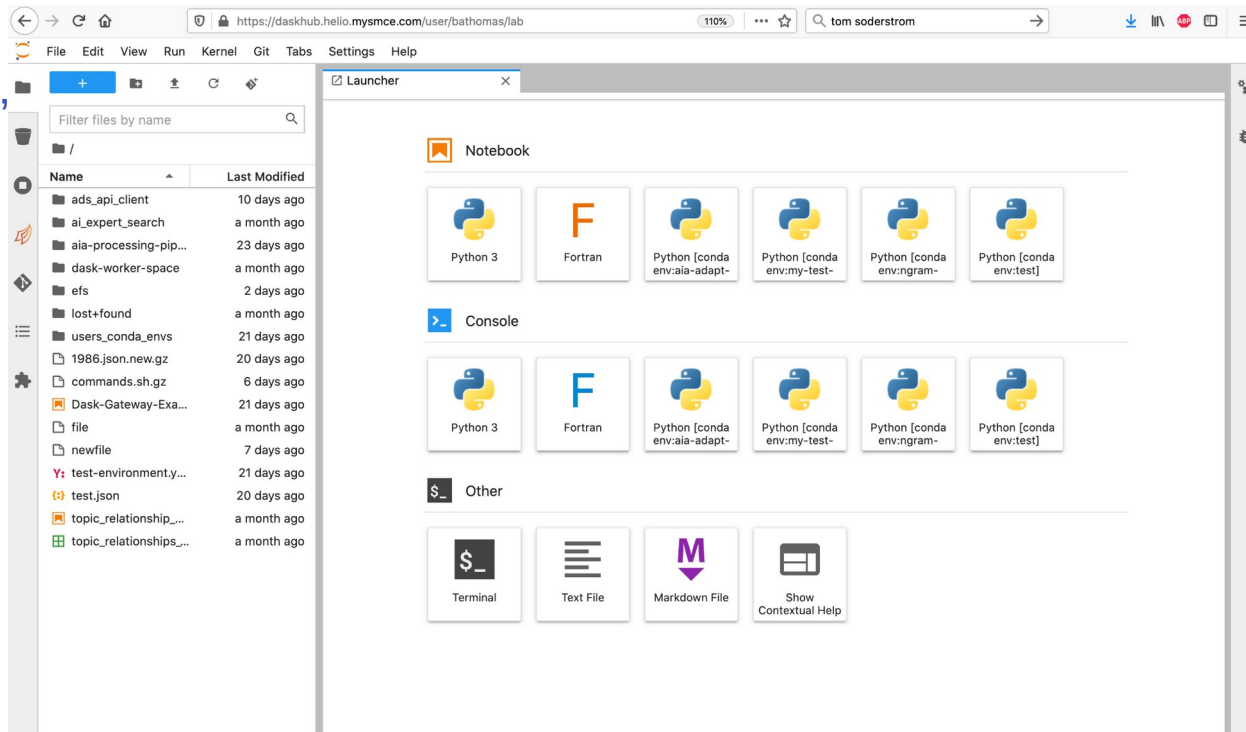
Daskhub + Notebooks

Heliophysics software environment which “just works”

Science Computing, integration with Containers

Deals with Large data volumes, “Bursts” into the cloud

Easy collaboration, strong sharing (data, code, results)



Shared Cloud Registry specification

A simple standard for any dataset so users can access it via API or directly. Summary:

- S3 disks have a 'catalog.json' describing their datasets
- Each dataset has <id>_YYYY.csv index files of its contents
- These indexes have the form of "Time, URI, filesize" (plus optional metadata)

```
[6]: file_registry1 = fr.request_file_registry(frID, start_date=start, stop_date=stop_date)
# And convert that richer data to a list of files to process
filelist = file_registry1['key'].to_list()
```

```
[7]: file_registry1
```

```
[7]:
```

	startDate	key	filesize
0	2020-02-17 00:00:00	s3://gov-nasa-hdrldata1/sdo/aia/20200217/0094...	13910400
1	2020-02-17 00:04:00	s3://gov-nasa-hdrldata1/sdo/aia/20200217/0094...	13910400
2	2020-02-17 00:08:00	s3://gov-nasa-hdrldata1/sdo/aia/20200217/0094...	13910400
3	2020-02-17 00:12:00	s3://gov-nasa-hdrldata1/sdo/aia/20200217/0094...	13910400


file registry
is very HAPI-like

time, key (s3 or other), filesize

HelioCloud - Overview : Capabilities for Researchers

Several capabilities available to researchers

- **Daskhub**
 - Notebooking & Terminal in Browser
 - 'Bursts' into the Cloud
 - Python / IDL / Fortran / MatLab
- **Portal Dashboard for VMs**
 - Ssh access
 - Static (fixed) IP addresses
 - Non-nasa hostnames
- **S3 Storage**
 - Public and Private access



*Available for GSFC (670) researchers
and their collaborators (including
international).*

FAST onboarding.

135+ Users in NASA instance of HelioCloud!

Future Progress:

- **High Level Data**

- SPDF : ALL CDAWeb (level 2+) : ~350 TB
- SDAC: ALL non-SDO (level 1+) : ~350 TB
- SDO : Selected AIA/HMI (level “1.5”): ~ 1 PB
- Contributed : AI/ML, other data : as space allows
- ~300 TB SDO data are public
- SPDF data to begin moving to public this semester
- SDAC data to begin moving to public this semester
- Contributed dataset: SDMOML (FDL data product) will be public ~week
- Multiuser shareable notebook in progress (Rebecca Ringuette)

- **Services**

- File registry : in development
- Science search : planned
- Better S3 Bucket browsing (Daskhub)
- User Identities (Daskhub)
- Improved Fortran Support
- Support for S3 Team buckets
- Better Financial Reporting
- ???

HelioCloud & PyHC:

Cloud environment for big data, open science, and collaboration

- We have Shared Interests!
- Make all of the software work well together
- Easy for users to access and use software
- Enabling an 'on-ramp' for new scientists and early careers
- Support existing 'entrenched' scientists' workflows
- Support Open Science
- Tackle big problems that were previously infeasible

Contact: Brian Thomas / HDRL
heliocloud.org

HDRL : What is it?

