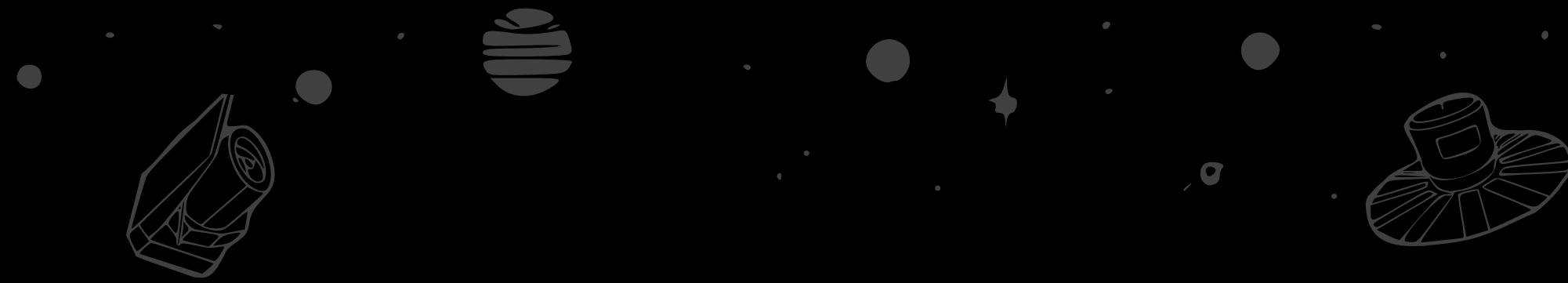


The SPACIOUS Platform

Brendan O’Brien

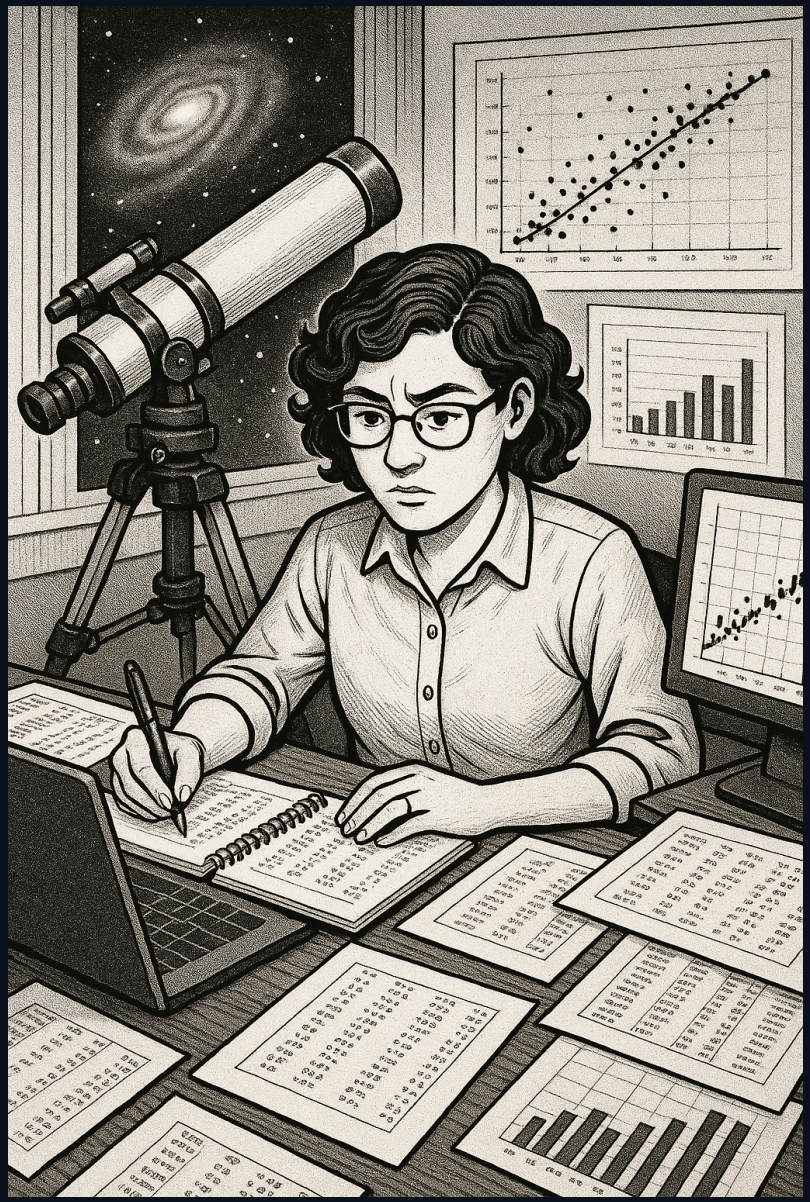
Royal Observatory of Edinburgh

brendan.obrien@roe.ac.uk

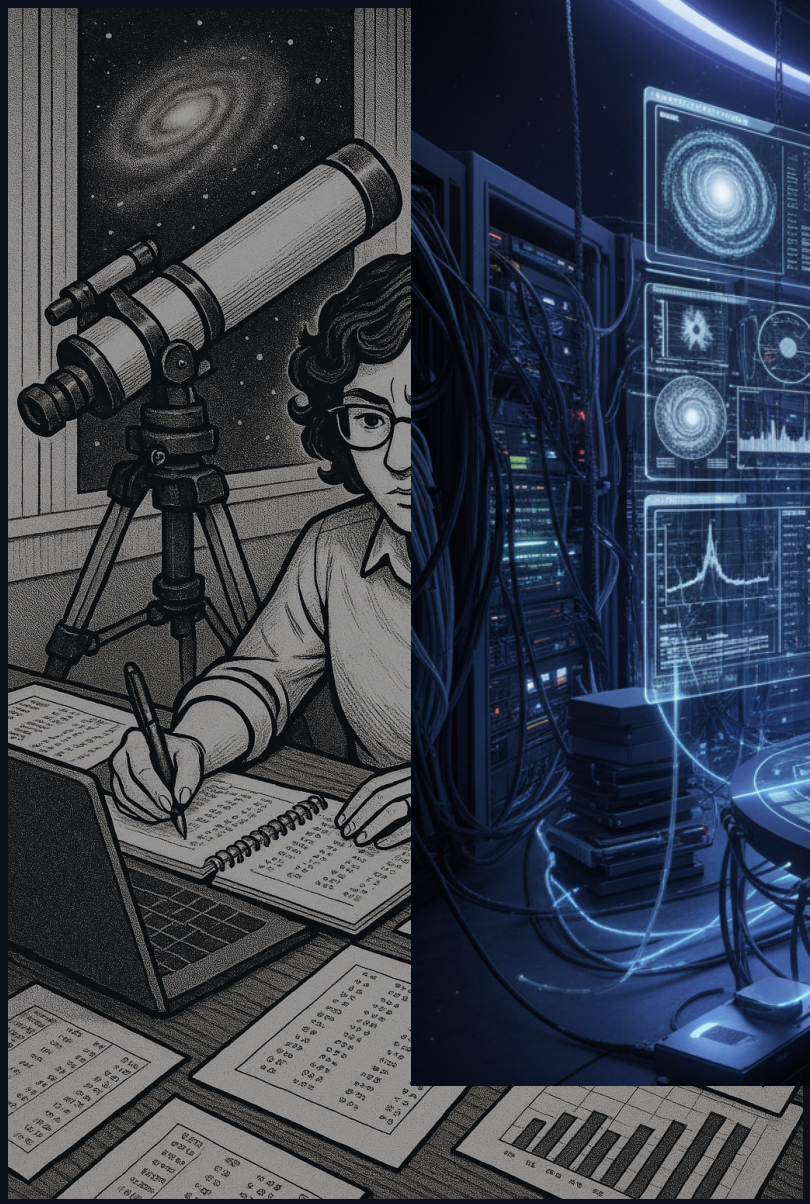
THE CHALLENGE

Modern astronomy is data intensive ...

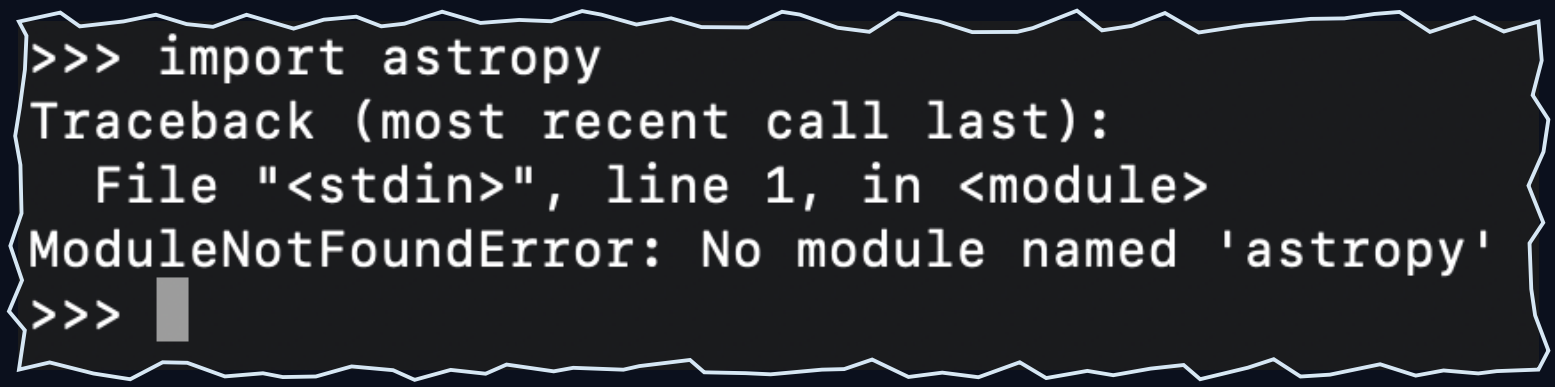
Modern astronomy is data intensive ...



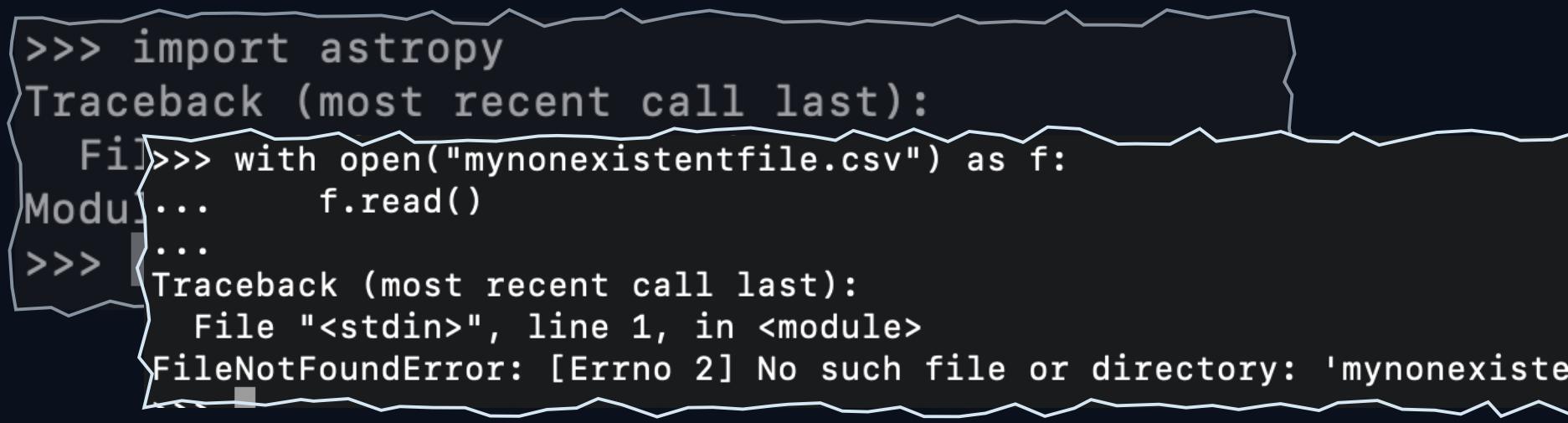
Modern astronomy is data intensive ...



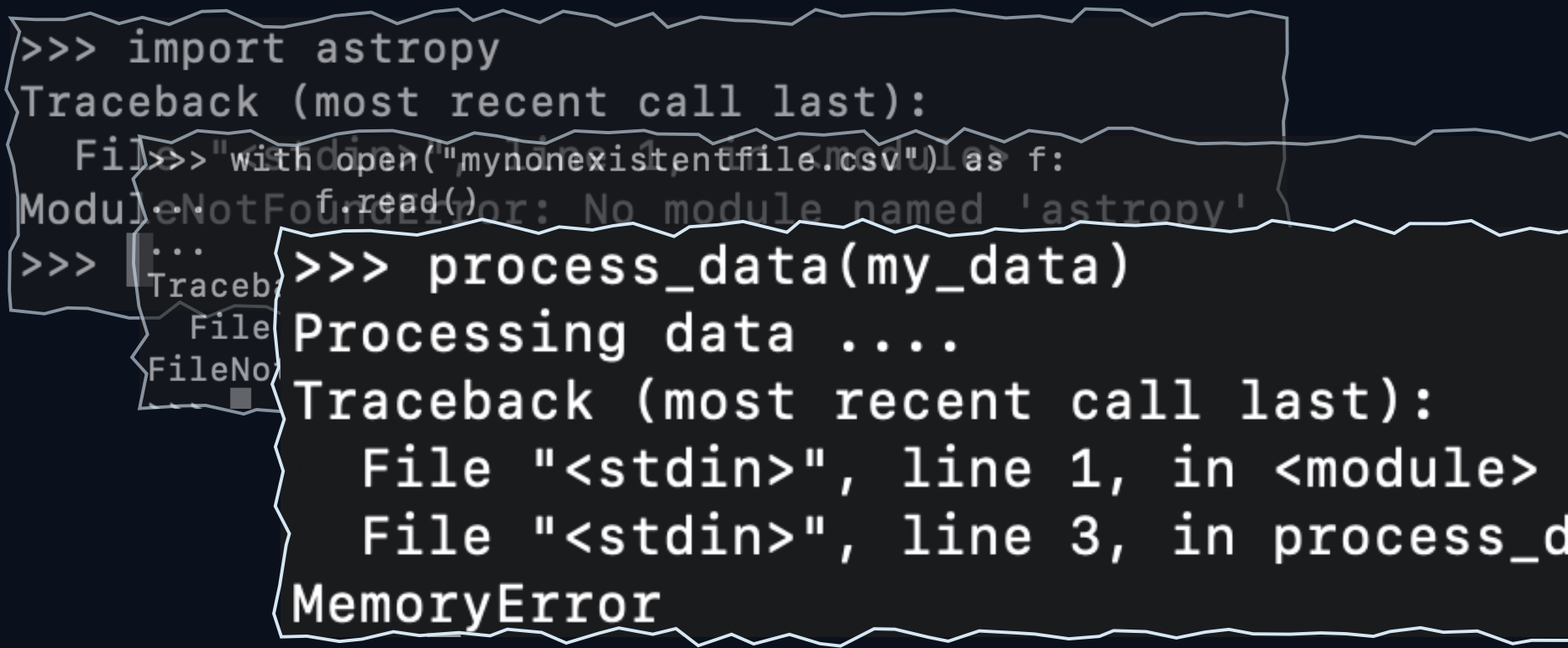
Fragmented tools $\rightarrow$ barriers to utilisation ...



Fragmented tools $\rightarrow$ barriers to utilisation ...



Fragmented tools → barriers to utilisation ...

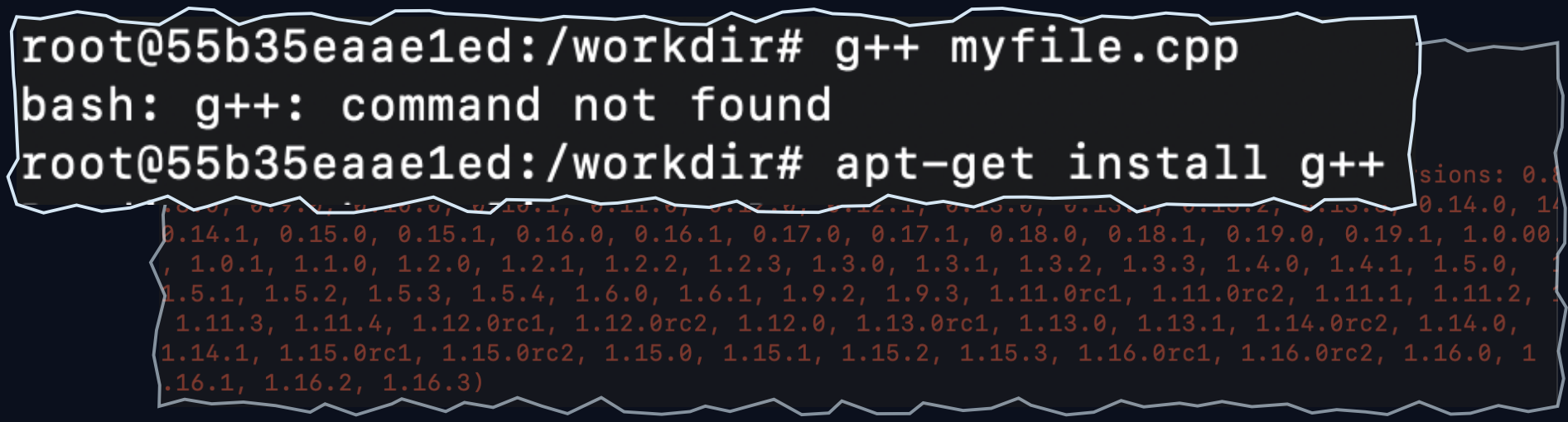


Installation & system challenges ...

Installation & system challenges ...

```
pip install -r requirements.txt
Collecting astropy==7.2.0 (from -r requirements.txt (line 1))
  Downloading astropy-7.2.0-cp311-abi3-macosx_11_0_arm64.whl.metadata (10 kB)
ERROR: Ignored the following yanked versions: 1.11.0, 1.14.0rc1
ERROR: Could not find a version that satisfies the requirement scipy==1.8.0 (from versions: 0.
8.0, 0.9.0, 0.10.0, 0.10.1, 0.11.0, 0.12.0, 0.12.1, 0.13.0, 0.13.1, 0.13.2, 0.13.3, 0.14.0, 1
0.14.1, 0.15.0, 0.15.1, 0.16.0, 0.16.1, 0.17.0, 0.17.1, 0.18.0, 0.18.1, 0.19.0, 0.19.1, 1.0.00
, 1.0.1, 1.1.0, 1.2.0, 1.2.1, 1.2.2, 1.2.3, 1.3.0, 1.3.1, 1.3.2, 1.3.3, 1.4.0, 1.4.1, 1.5.0,
1.5.1, 1.5.2, 1.5.3, 1.5.4, 1.6.0, 1.6.1, 1.9.2, 1.9.3, 1.11.0rc1, 1.11.0rc2, 1.11.1, 1.11.2,
 1.11.3, 1.11.4, 1.12.0rc1, 1.12.0rc2, 1.12.0, 1.13.0rc1, 1.13.0, 1.13.1, 1.14.0rc2, 1.14.0,
1.14.1, 1.15.0rc1, 1.15.0rc2, 1.15.0, 1.15.1, 1.15.2, 1.15.3, 1.16.0rc1, 1.16.0rc2, 1.16.0, 1
.16.1, 1.16.2, 1.16.3)
```

Installation & system challenges ...



Local machines can’t handle the TB + scale ...

Scientists should be scientists ...

Scientists should be scientists ...

Scientists should be scientists ... not cluster admins!!!


Data to code ...


Code to Data ...

THE GOAL

Science PlAtform Cloud Infrastructure for Outsize Usage Scenarios

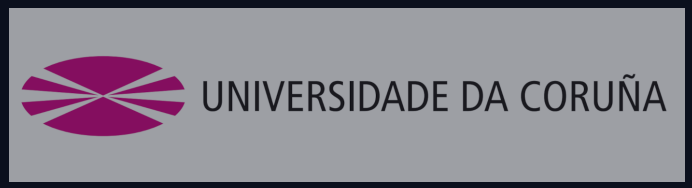
Funded by the European Union

SPACIOUS is a Horizon Europe HORIZON-CL4-2023-SPACE-01-71 project funded under grant agreement no. 101135205

https://spacious.ub.edu/

“Aims to become a turning point for exploiting scientific data from space missions more efficiently, through a new computational framework in astrophysics based on big data and data mining technologies”

Collaboration ...

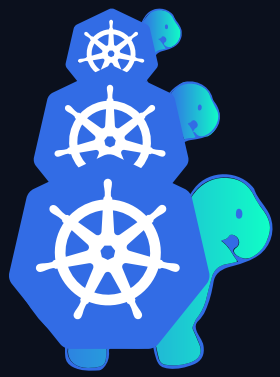


The New Computational Framework ...

https://astro-flow.com

KUBERNETES

Cloud Native ...



Cloud Native ...

openstack™

Cloud Native ...

Google Cloud

openstack™

Cloud Native ...

MINIO

LONGHORN

Google Cloud

JuiceFS

Cloud Native ...


OBSERVABILITY

Observability ...

Observability ...

Observability ...

Observability ...

filename	/var/log/pods/ingress-nginx_ingress-ngin
instance	ingress-nginx
job	ingress-nginx/ingress-nginx
method	
namespace	ingress-nginx
node_name	bob-test-cluster-bucubogk2bqm-default-wo
path	
pod	ingress-nginx-controller-54b765b7b-td7jm
remote_addr	
remote_user	
req_body	
req_content_type	
req_id	25d7a6aad75ad756dc6ddc5ef3718cab
request_id	25d7a6aad75ad756dc6ddc5ef3718cab


Security ...

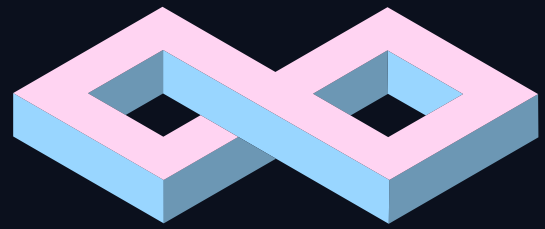
Security ...

Security ...

Security ...

Security ...

Security ...



aqua
kube-bench


ALERTING

Alerting ...

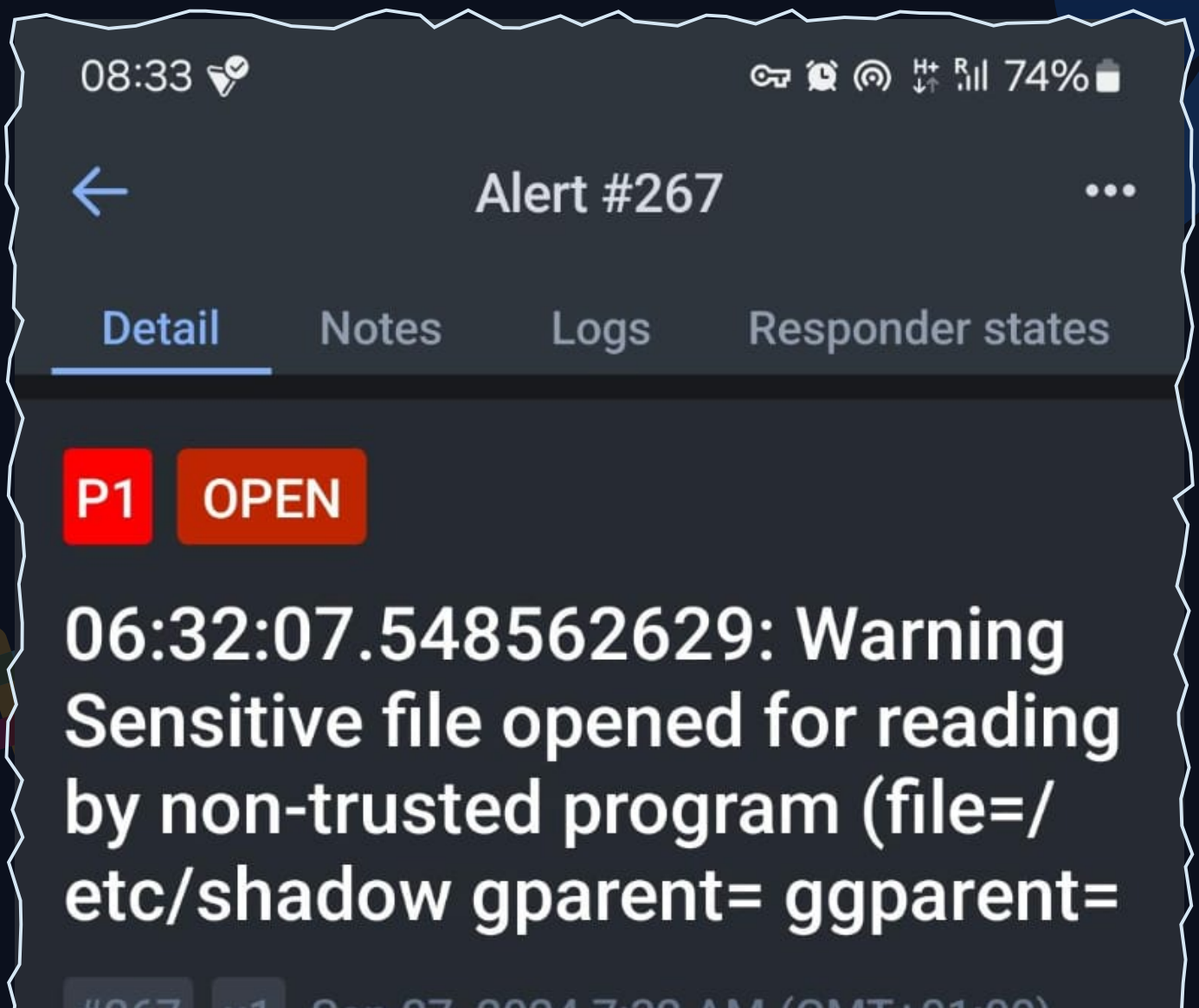
Alerting ...

AstroFlow Webhook APP 07:32

06:32:07.548562629: Warning Sensitive file opened for reading by non-trusted program (file=/etc/shadow gparent=<NA> ggparent=<NA> gggparent=<NA> evt_type=openat user=root user_uid=0 user_loginuid=-1 process=cat proc_exepath=/usr/bin/cat parent=systemd command=cat /etc/shadow terminal=34816 container_id=117742f17879 container_image=docker.io/library/nginx container_image_tag=latest container_name=nginx k8s_ns=nginx k8s_pod_name=nginx-deployment-968f5bf77-p44zx)

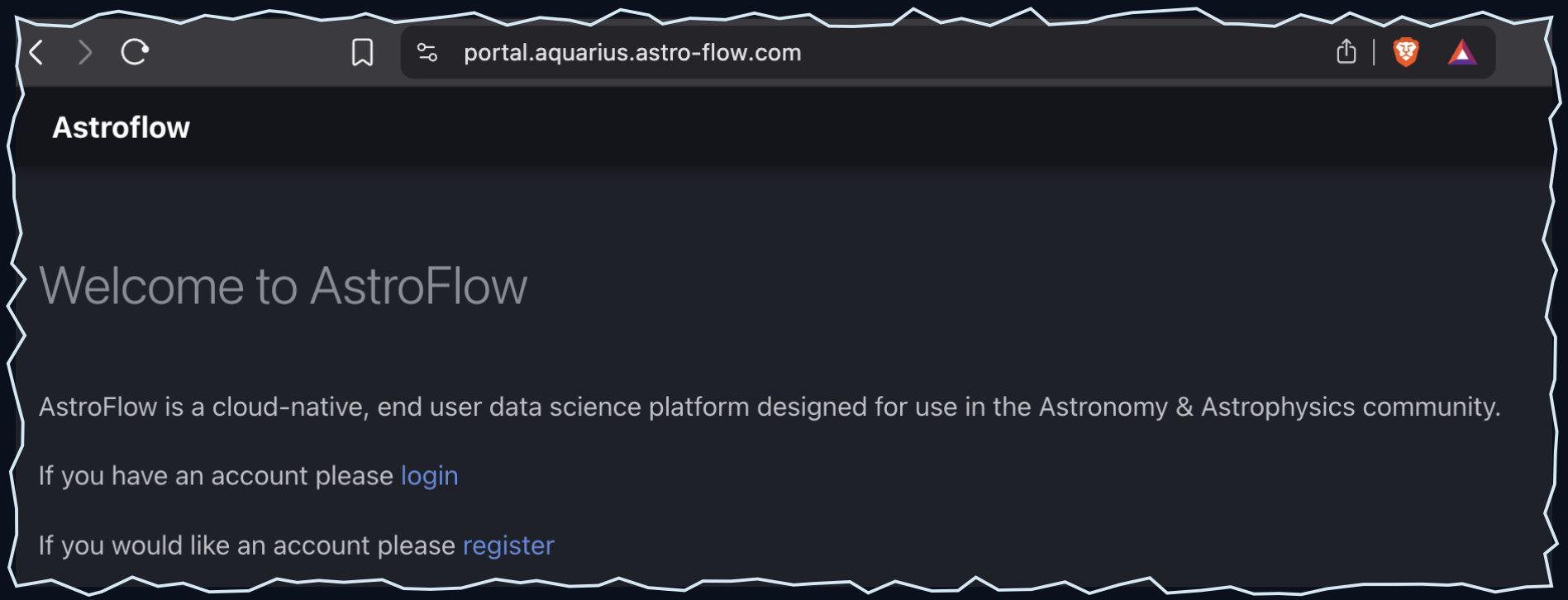
rule Read sensitive file untrusted	**priority** Warning
source syscall	**hostname** bob-test-cluster-2-aja2w23r6tyb-default-worker-rl7fg-nbfb7
tags T1555, container, filesystem, host, maturity_stable, mitre_credential_access	**container.id** 117742f17879
container.image.repository docker.io/library/nginx	**container.image.tag** latest
container.name nginx	**evt.type** openat
fd.name /etc/shadow	**k8s.ns.name** nginx
k8s.pod.name nginx-deployment-968f5bf77-p44zx	**proc.cmdline** cat /etc/shadow

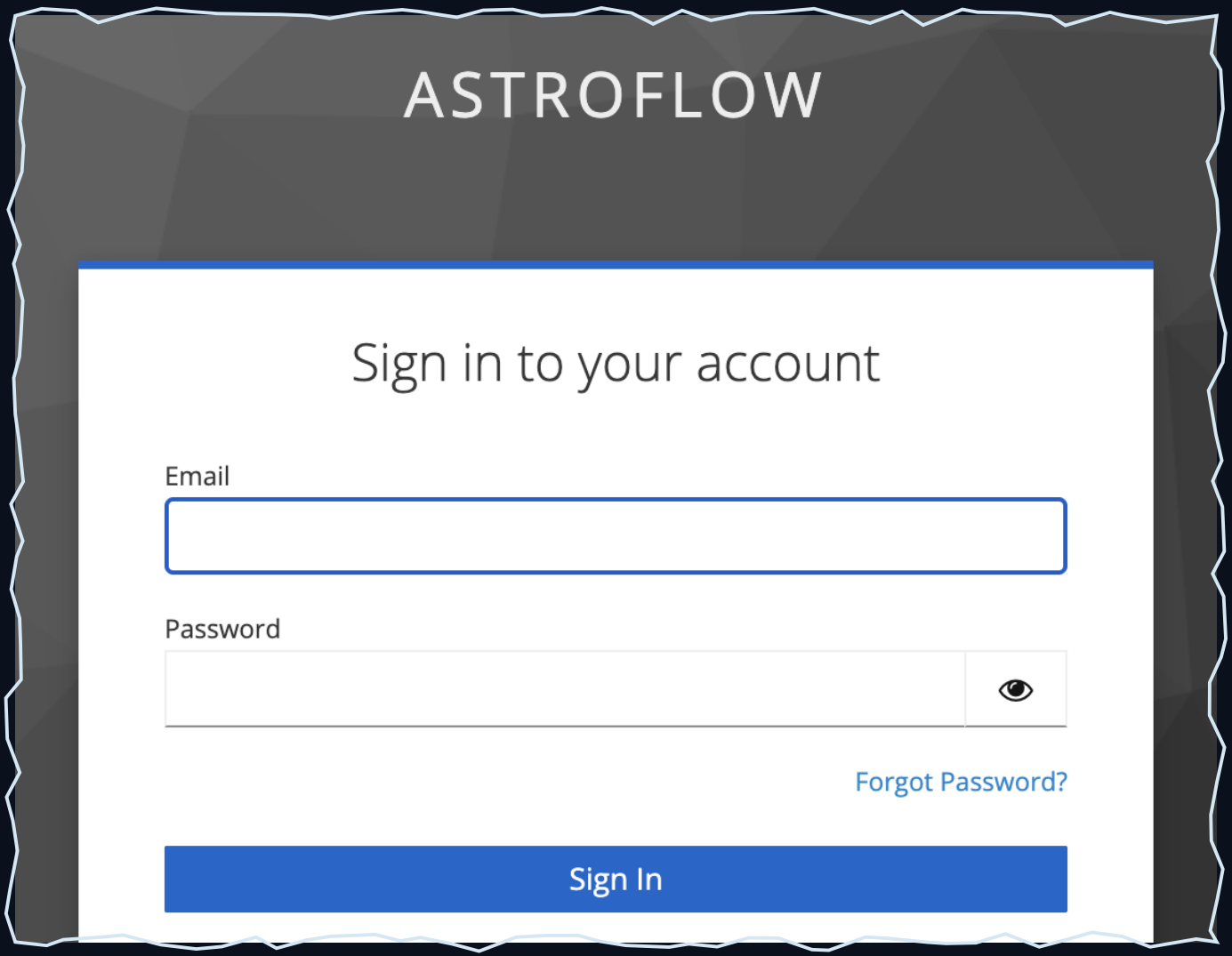
Alerting ...


08:33
74%
Alert #267
Detail
Notes
Logs
Responder states
P1
OPEN
06:32:07.548562629: Warning Sensitive file opened for reading by non-trusted program (file=/etc/shadow gparent= ggparent=

End User Feature Walk Through ...

The Portal ...


portal.aquarius.astro-flow.com
Astroflow
Welcome to AstroFlow
AstroFlow is a cloud-native, end user data science platform designed for use in the Astronomy & Astrophysics community.
If you have an account please login
If you would like an account please register


ASTROFLOW
Sign in to your account
Email
Password
Forgot Password?
Sign In

My Profile

My Details

- **Name:** Brendan O'Brien
- **Email:** brendan.obrien@roe.ac.uk
- **Unique UserID (UUID):** 247d4c07-a759-4e0f-b913-8838237e7c02 (this will be used by the system to badge your sessions)

Update your details

Please note that updated details will only reflect here after triggering a token refresh (logout and login)

My Services

- **Portal**

  The page you are currently on, where you can check your details and entitlements

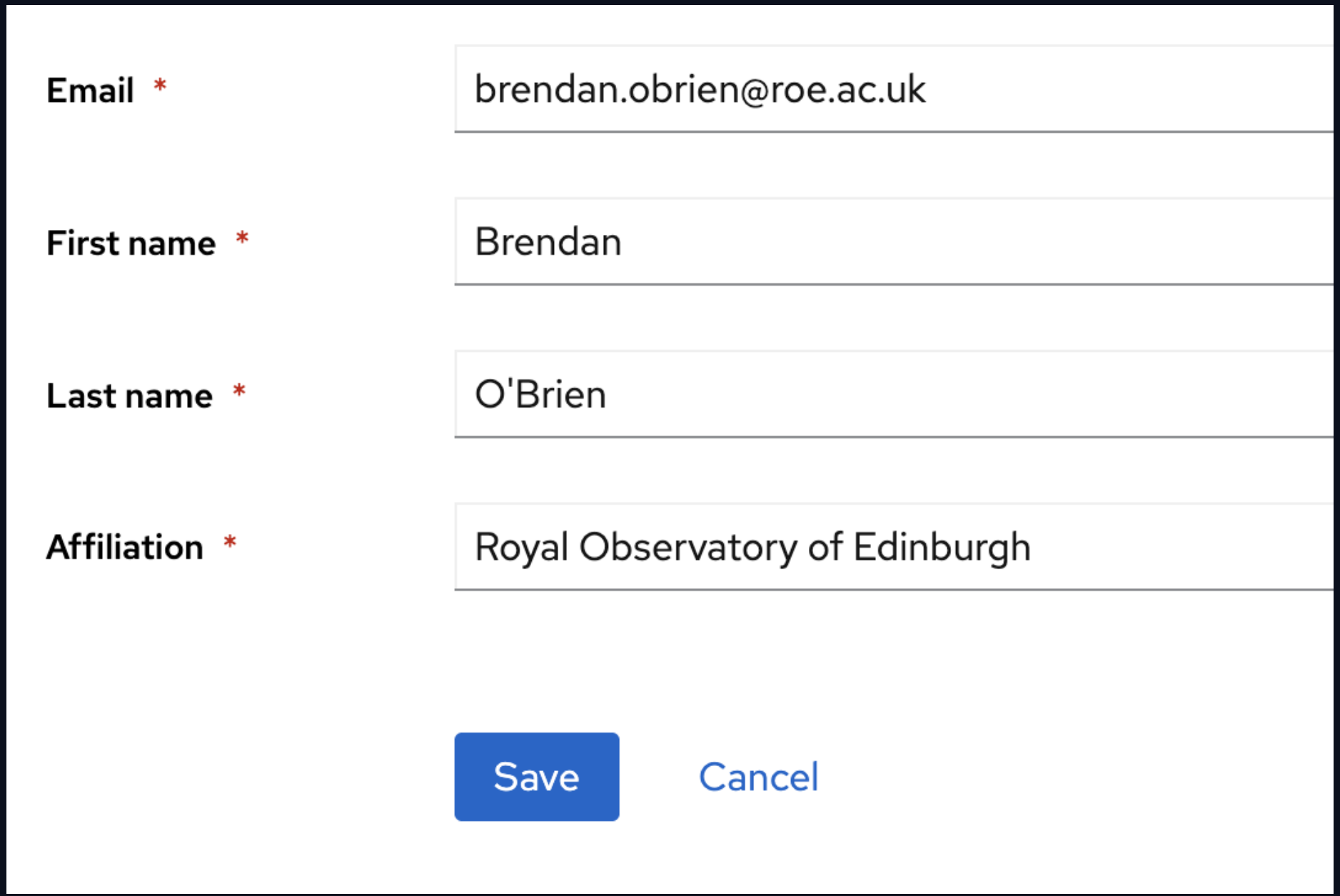
- **Documentation**

  Access to the AstroFlow documentation here

- **Jupyter Hub**

  Access to jupyterhub where you can make use of pre-installed packages and spark and dask clusters

  Access the hub here


Email *
brendan.obrien@roe.ac.uk
First name *
Brendan
Last name *
O'Brien
Affiliation *
Royal Observatory of Edinburgh
Save
Cancel

Signing in

Configure ways to sign in.

Basic authentication

Password

Sign in by entering your password.

My password **Created** October 30, 2024 at 6:19 PM. Update

Two-factor authentication

Authenticator application

Set up Authenticator application

Enter a verification code from authenticator application.

Authenticator application is not set up.

Documentation ...

Promtail

Promtail is a log collector that works alongside Loki, designed to gather logs from various sources and ship them to Loki for storage and querying. Promtail is lightweight and tailors itself to kubernetes environments by automatically discovering running pods and their associated logs using kubernetes labels, ensuring logs are efficiently tagged and indexed.

Promtail integrates with systemd journals, log files, and other logging infrastructures, and uses a configuration-based approach to define how logs should be collected and processed. It also supports transforming logs using pipelines before sending them to Loki, ensuring logs are enriched and filtered as needed.

Core functionality

Cert Manager

cert-manager automates automates the issue, management and renewal of TLS certificates. It integrates with Let's Encrypt and other certificate providers.

Table of contents

AstroFlow Architecture

The overall cluster architecture is shown below.

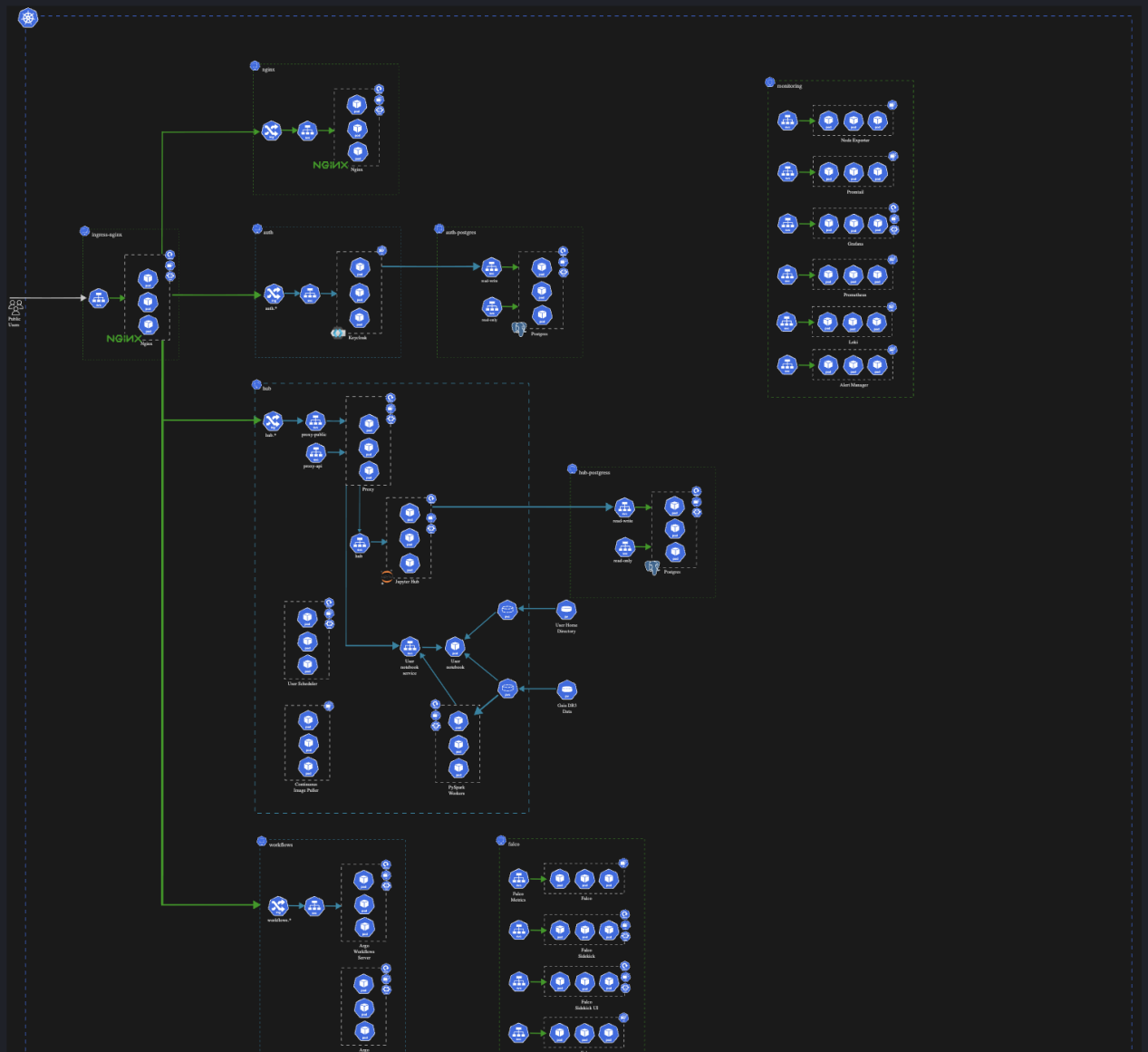


Table of contents

Developer Guide

The AstroFlow application level code lives here and the project board lives here

Development workflow

AstroFlow uses a GitFlow approach to development.

Installation Example (Openstack)

> **Warning**
>
> If you're reading this and intending to deploy to arcus, with the ceph storage mounts, then it would be better to clone from arcus demo rather than somerville demo as per the guide here. As the storage mounts will be the same etc. It will be a closer starting point.

> **Note**
>
> The full documentation is here, this page is a hands-on example run through.
>
> It is designed to be used in conjunction with the full installation instructions, which include explanations

Cluster

For this example we are using an openstack cluster with 1 master node and 3 worker nodes, it's empty, aside from some utilities that were pre-installed on it, as part of the openstack/magnum/coe cluster provisioning process. You can assume it's empty.

Just testing the config with the following command to ensure we can see some nodes.

```
kubectl get nodes
NAME                                                STATUS   ROLES           A(
bob-test-demo-ku65xockwdzj-control-plane-76m9q      Ready    control-plane   1(
bob-test-demo-ku65xockwdzj-default-worker-9bst9-84q5n   Ready    <none>      6r
bob-test-demo-ku65xockwdzj-default-worker-9bst9-cdklx   Ready    <none>      5r
bob-test-demo-ku65xockwdzj-default-worker-9bst9-sbhkw   Ready    <none>      6r
```

Table of contents

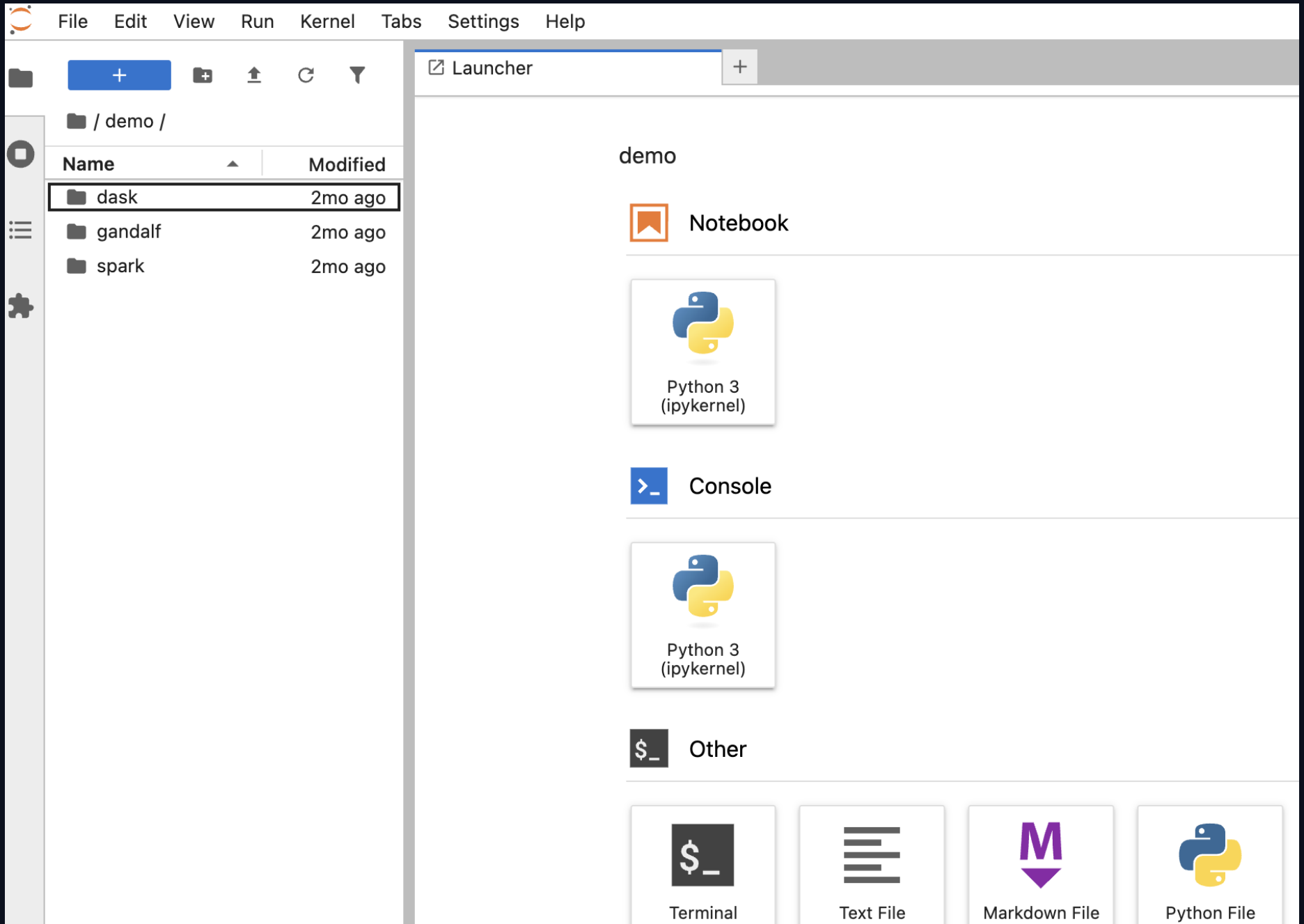
The hub ...

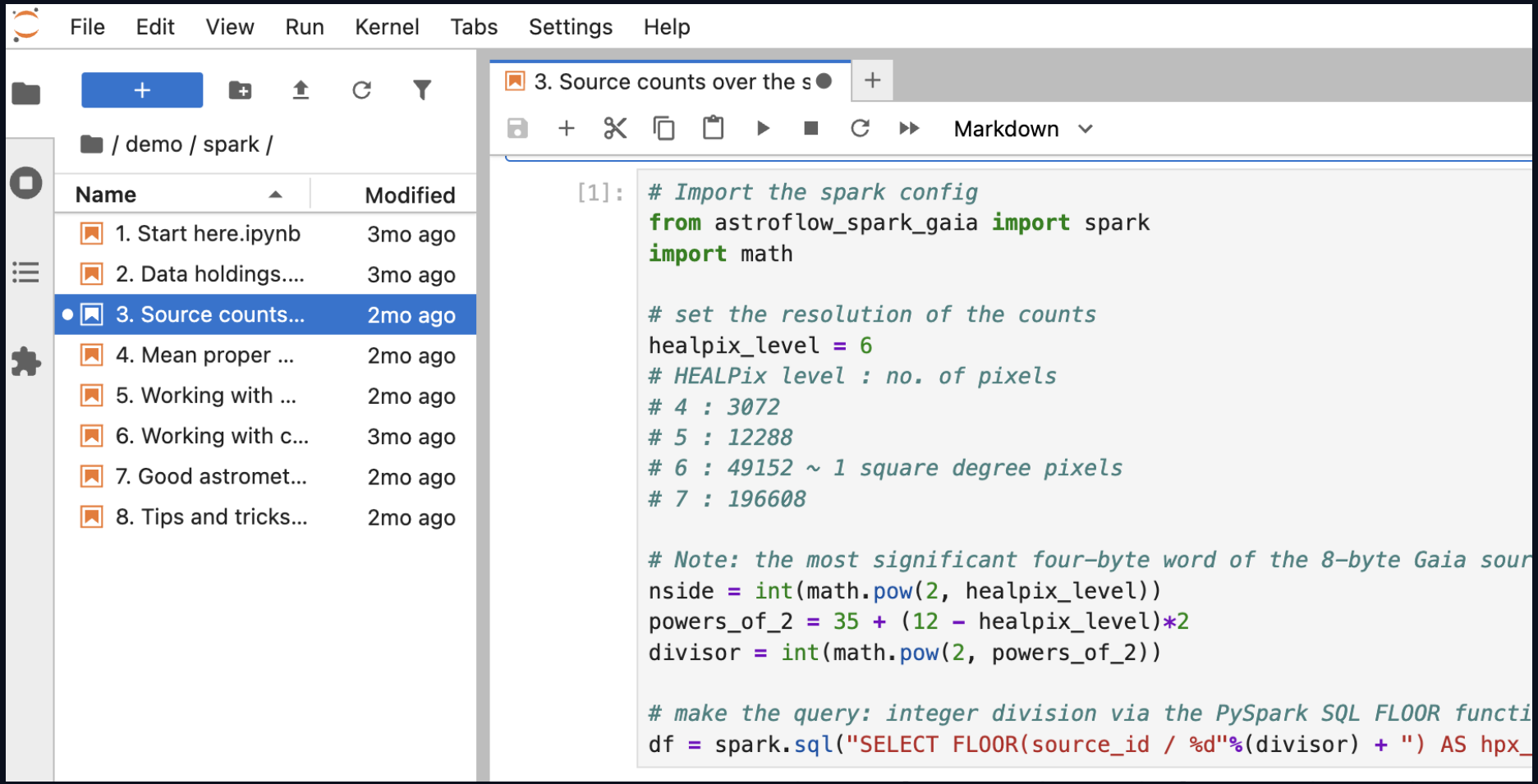
Server Options

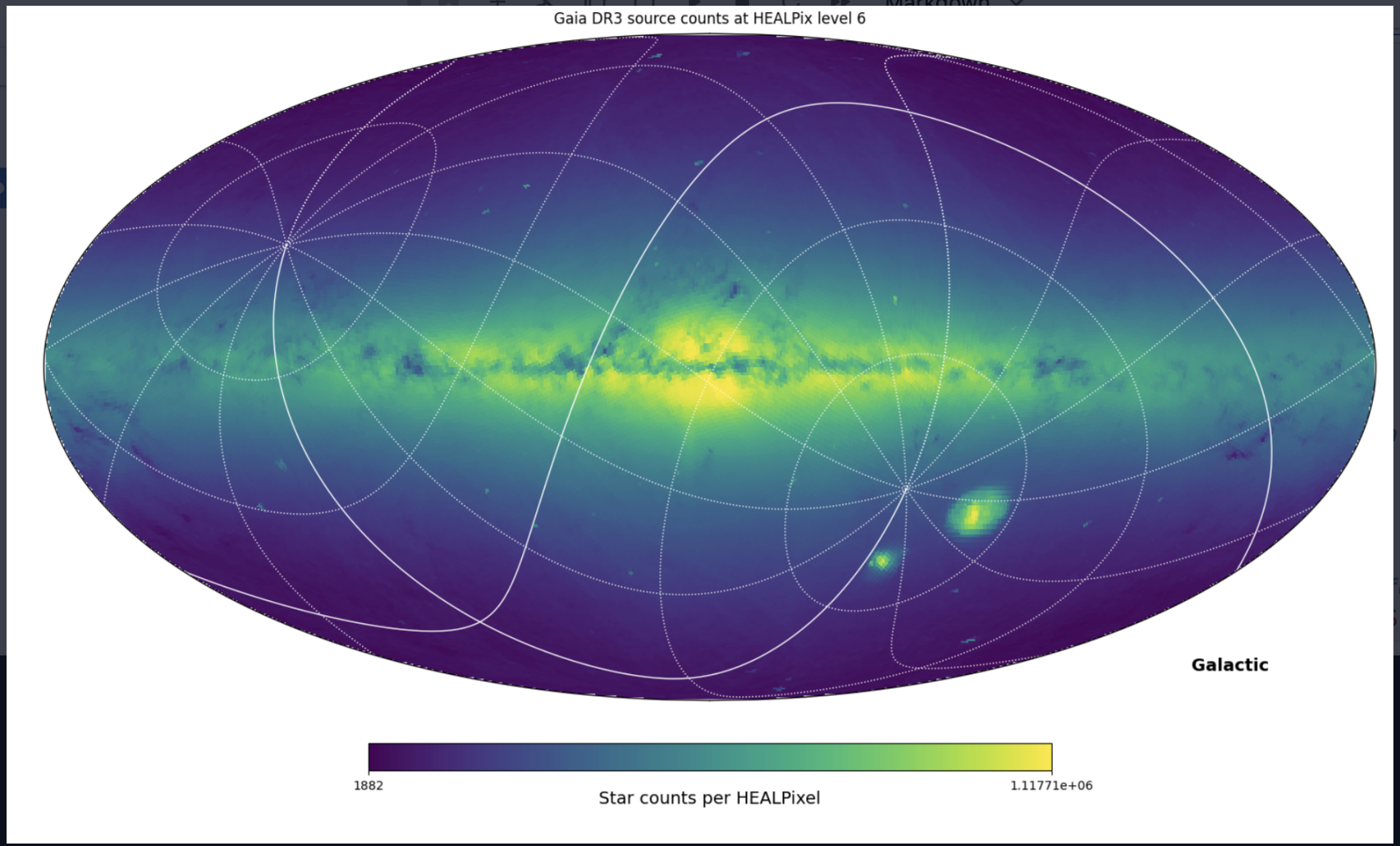
Gaia DR3 Environment

In order to use pre-configured Apache Spark and Dask with Gaia DR3 data, use this environment

Start


File
Edit
View
Run
Kernel
Tabs
Settings
Help
Launcher
/ demo /
Name
Modified
dask
2mo ago
gandalf
2mo ago
spark
2mo ago
demo
Notebook
Python 3 (ipykernel)
Console
Python 3 (ipykernel)
Other
Terminal
Text File
Markdown File
Python File


File Edit View Run Kernel Tabs Settings Help
/ demo / spark /
Name
Modified
1. Start here.ipynb 3mo ago
2. Data holdings.... 3mo ago
3. Source counts... 2mo ago
4. Mean proper ... 2mo ago
5. Working with ... 2mo ago
6. Working with c... 3mo ago
7. Good astromet... 2mo ago
8. Tips and tricks... 2mo ago
3. Source counts over the s
Markdown
[1]:
# Import the spark config
from astroflow_spark_gaia import spark
import math
# set the resolution of the counts
healpix_level = 6
# HEALPix level : no. of pixels
# 4 : 3072
# 5 : 12288
# 6 : 49152 ~ 1 square degree pixels
# 7 : 196608
# Note: the most significant four-byte word of the 8-byte Gaia sour
nside = int(math.pow(2, healpix_level))
powers_of_2 = 35 + (12 - healpix_level)*2
divisor = int(math.pow(2, powers_of_2))
# make the query: integer division via the PySpark SQL FLOOR functi
df = spark.sql("SELECT FLOOR(source_id / %d"%(divisor) + ") AS hpx_


Gaia DR3 source counts at HEALPix level 6
Galactic
1882
1.11771e+06
Star counts per HEALPixel


APACHE
Spark
TM

[1]:
```
# Import the spark config
from astroflow_spark_gaia import spark
```

```
Beginning the initilization of a spark cluster with 10 pods, 5 cores/pod and 7g mem/pod
Completed initialisation
Setting up SparkSQL
Setting up dataset Gaia DR3 with 4 databases gaiadr3, gaiadr3ssd, gaiaedr3ssd, gaiaedr3
Read default database gaiadr3ssd from environment config
A spark cluster has been sucessfully set up, you can interact with it via the "spark" object
```

[1]:
```python
# Import the spark config
from astroflow_spark_gaia import spark
```

```
Beginning the initilization of a spark cluster with 10 pods, 5 cores/pod and 7g mem/pod
Completed initialisation
Setting up SparkSQL
Setting up dataset Gaia DR3 with 4 databases gaiadr3, gaiadr3ssd, gaiaedr3ssd, gaiaedr3
Read default database gaiadr3ssd from environment config
A spark cluster has been sucessfully set up, you can interact with it via the "spark" object
```

[2]:
```python
df = spark.sql("SELECT COUNT(*) FROM gaia_source")
df.collect()
```

[2]:
```
[Row(count(1)=1811709771)]
```


dask

[13]:

```
# Import the dask client
from dask.distributed import Client
import dask.dataframe as dd

c = Client('tcp://simple-scheduler.dask-operator.svc.cluster.local:8786')
print(c)
```

```
<Client: 'tcp://172.19.205.22:8786' processes=10 threads=50, memory=700.00 GiB>
```

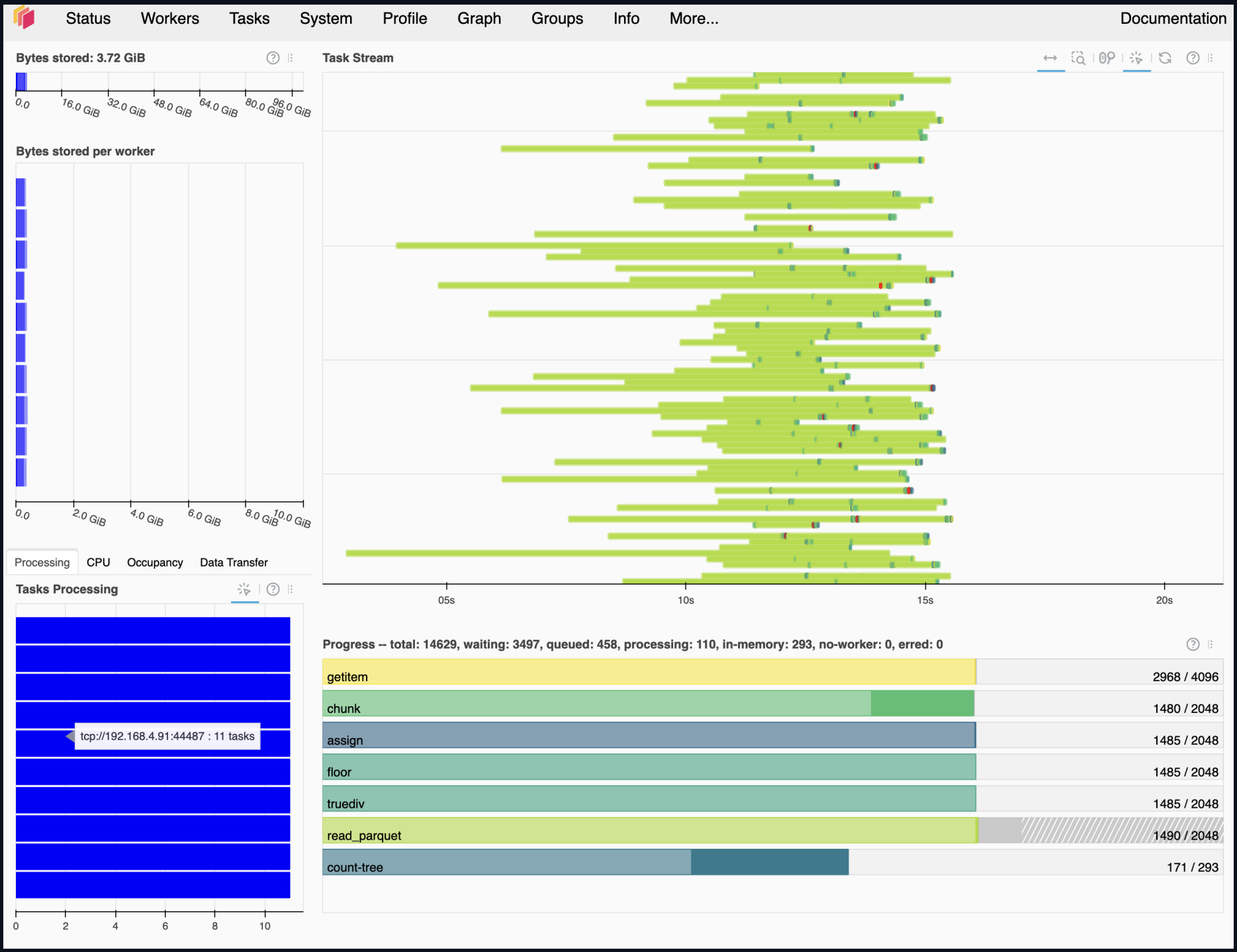
[13]:
```python
# Import the dask client
from dask.distributed import Client
import dask.dataframe as dd

c = Client('tcp://simple-scheduler.dask-operator.svc.cluster.local:8786')
print(c)
```

```
<Client: 'tcp://172.19.205.22:8786' processes=10 threads=50, memory=700.00 GiB>
```

[15]:
```python
ddf = dd.read_parquet('/mnt/gaia-dr3-data-ssd-dask/GDR3_GAIA_SOURCE')
ddf.count().compute()
```

[15]:
```
solution_id                1811709771
designation                1811709771
source_id                  1811709771
random_index               1811709771
ref_epoch                  1811709771
                              ...
ag_gspphot_upper            470759263
ebpminrp_gspphot            470759263
ebpminrp_gspphot_lower      470759263
ebpminrp_gspphot_upper      470759263
libname_gspphot             470759263
Length: 152, dtype: int64
```


Status
Workers
Tasks
System
Profile
Graph
Groups
Info
More...
Documentation
Bytes stored: 3.72 GiB
0.0
16.0 GiB
32.0 GiB
48.0 GiB
64.0 GiB
80.0 GiB
96.0 GiB
Bytes stored per worker
0.0
2.0 GiB
4.0 GiB
6.0 GiB
8.0 GiB
10.0 GiB
Task Stream
05s
10s
15s
20s
Processing
CPU
Occupancy
Data Transfer
Tasks Processing
tcp://192.168.4.91:44487 : 11 tasks
0
2
4
6
8
10
Progress -- total: 14629, waiting: 3497, queued: 458, processing: 110, in-memory: 293, no-worker: 0, erred: 0
getitem
2968 / 4096
chunk
1480 / 2048
assign
1485 / 2048
floor
1485 / 2048
truediv
1485 / 2048
read_parquet
1490 / 2048
count-tree
171 / 293

What you get ...

What you get ...

Cross cluster access to spark and dask

What you get ...

Cross cluster access to spark and dask

30GB personal storage (backed up)

What you get ...

Cross cluster access to spark and dask

30GB personal storage (backed up)

Access to GAIA DR3 data

What you get ...

Cross cluster access to spark and dask

30GB personal storage (backed up)

Access to GAIA DR3 data

Pre-baked environments, optimised & managed by us

What you get ...

Cross cluster access to spark and dask
30GB personal storage (backed up)
Access to GAIA DR3 data
Pre-baked environments, optimised & managed by us
GAIA Visualisation & Guasom

What you get ...

Cross cluster access to spark and dask

30GB personal storage (backed up)

Access to GAIA DR3 data

Pre-baked environments, optimised & managed by us

GAIA Visualisation & Guasom

The ability to manage your own local installs in your environments


NEXT STEPS

What’s next ...

GAIA DR4

What’s next ...

GAIA DR4

Euclid tools

What’s next ...

GAIA DR4

Euclid tools

Batch mode processing

What’s next ...

GAIA DR4

Euclid tools

Batch mode processing

Feedback