CSE 4510/5310
Big Data

Instructor: Fitzroy Nembhard, Ph.D.

SaturnCloud

Setting up a Saturn Cloud Dask cluster, Loading Data into an S3 Bucket and Performing Operations on the Data

FLORIDA TECH
FLORIDA'S STEM UNIVERSITY®

Goals

- To set up a dask cluster on Saturn Cloud
- To Communicate with the dask cluster via Jupyter
- To access data in S3 via the dask cluster

Saturn Cloud Pricing

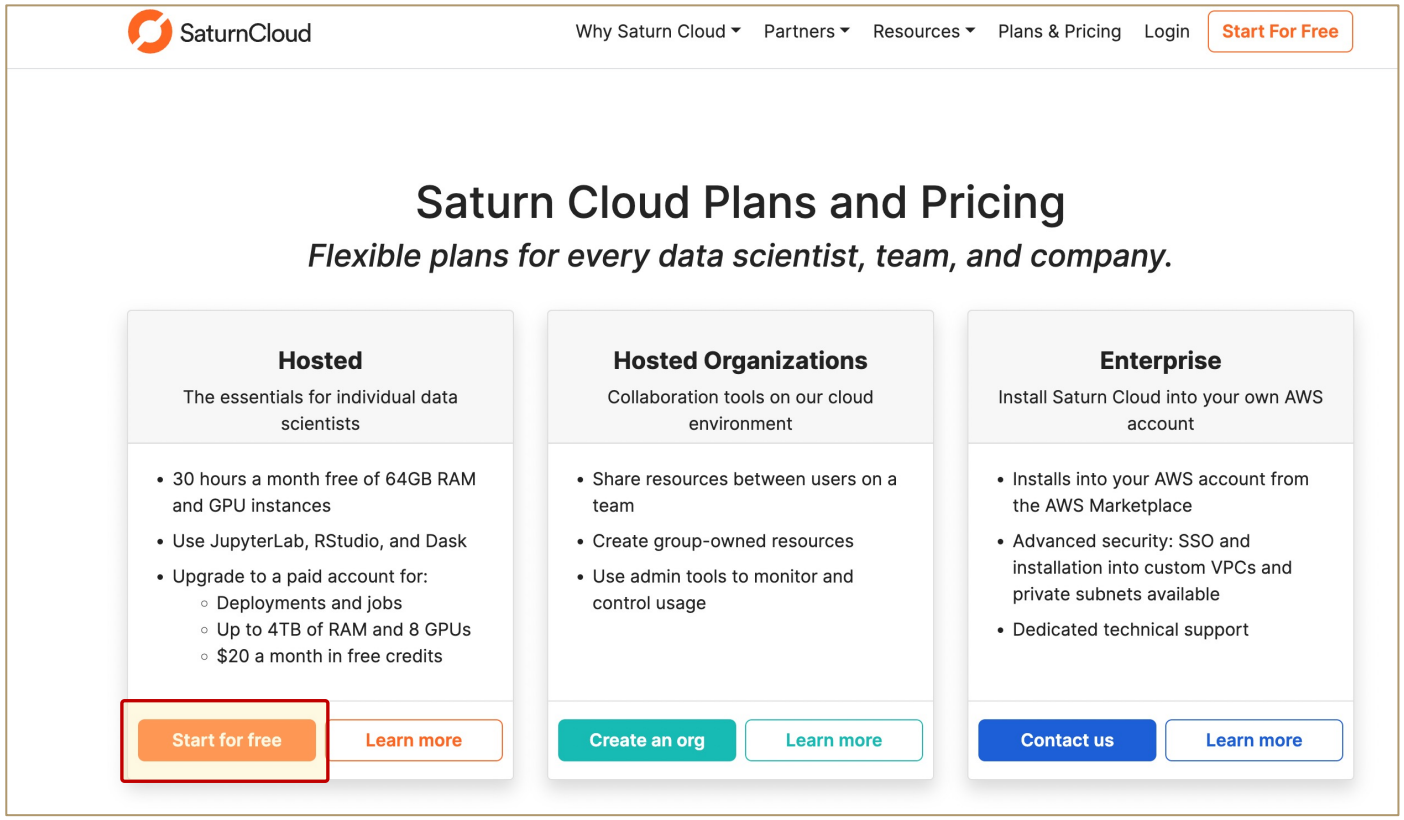
Click the following link to access Saturn Cloud.

https://saturncloud.io/plans/saturn_cloud_plans/

Then click Start for free.

You will have 30 hours worth of time on the server per month. To maximize your results, test your work locally using a pseudo dask cluster on your computer. Once you get things working, spin up your dask cluster and reproduce the results.

Save your work periodically to AWS S3 as parquet files. Stop your cluster when not in use. You may sign up with multiple email addresses to get more time.



FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Create an Account

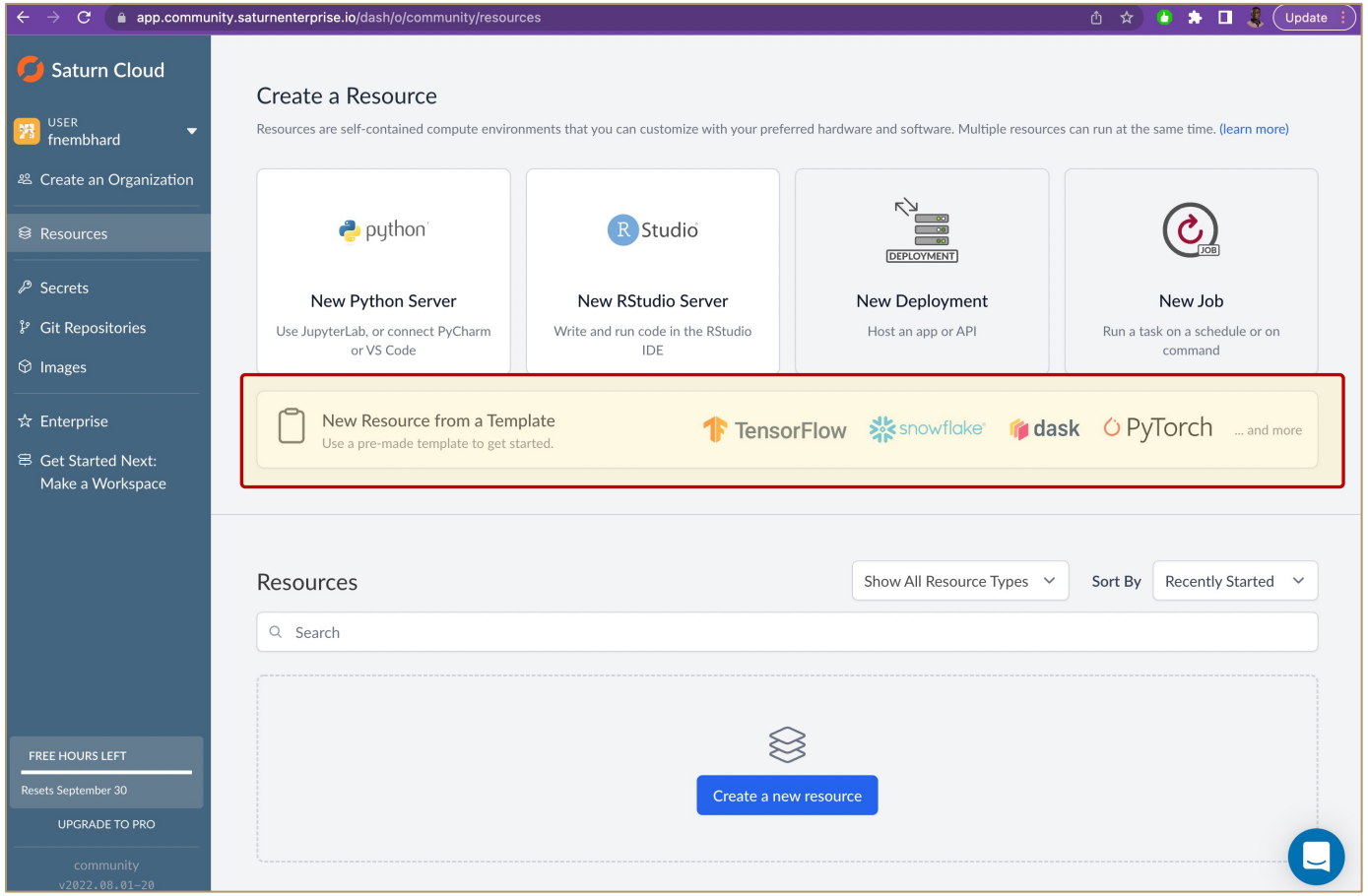
Choose your favorite option to create an account.

FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Sign In

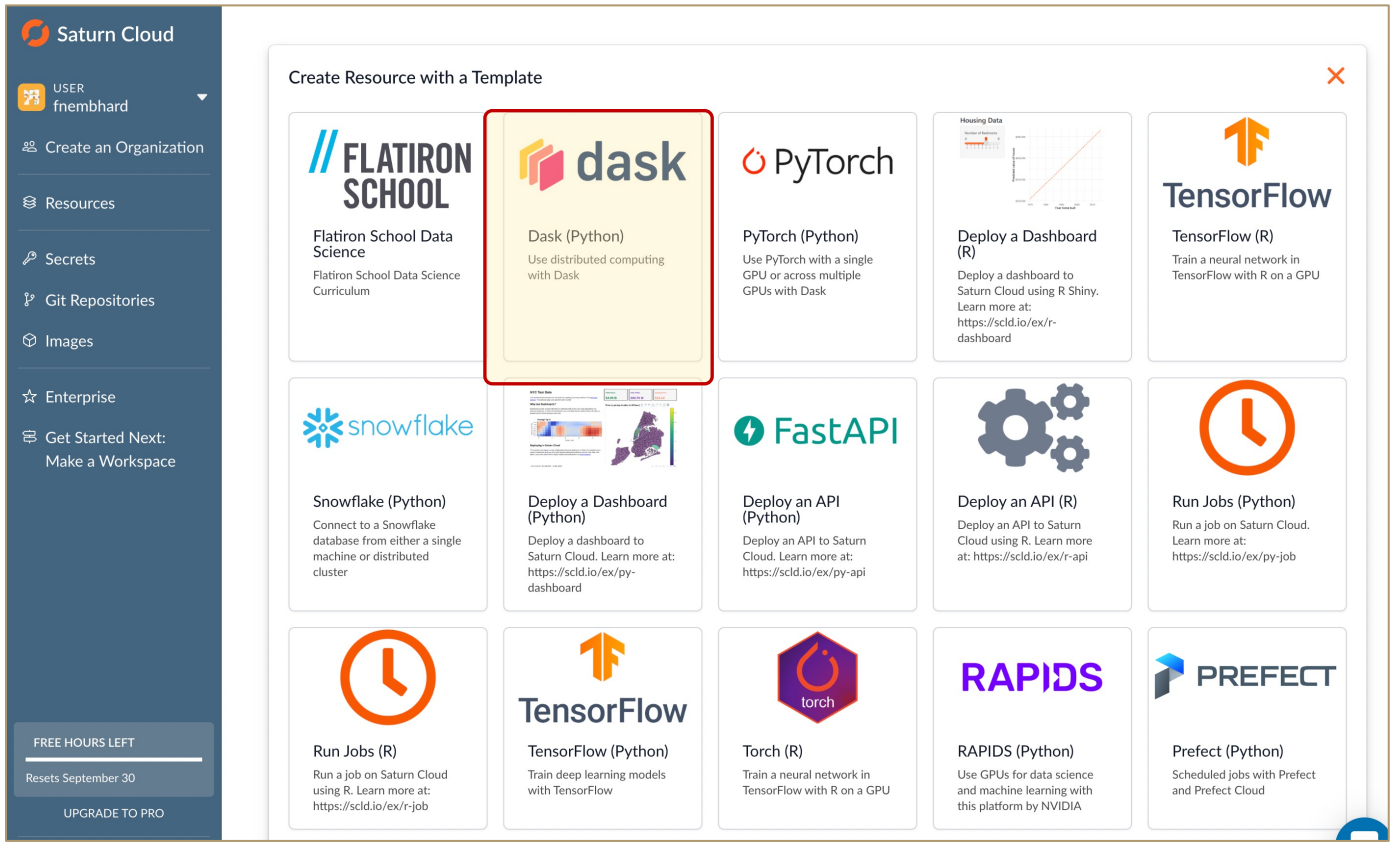
After logging in, you should see the following screen.

Click New Resource from a Template



Select Dask

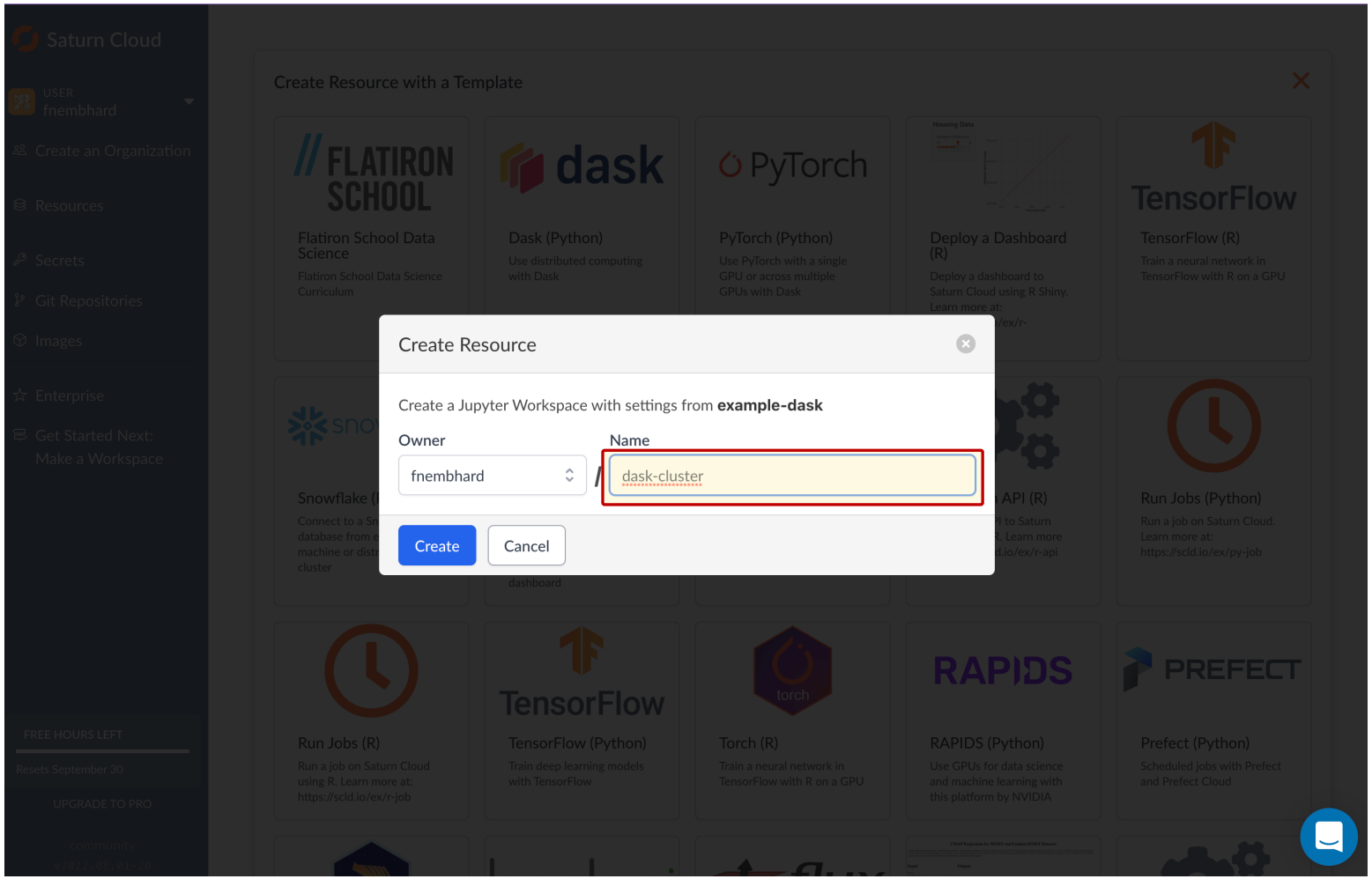
Click dask to access a resource template



FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

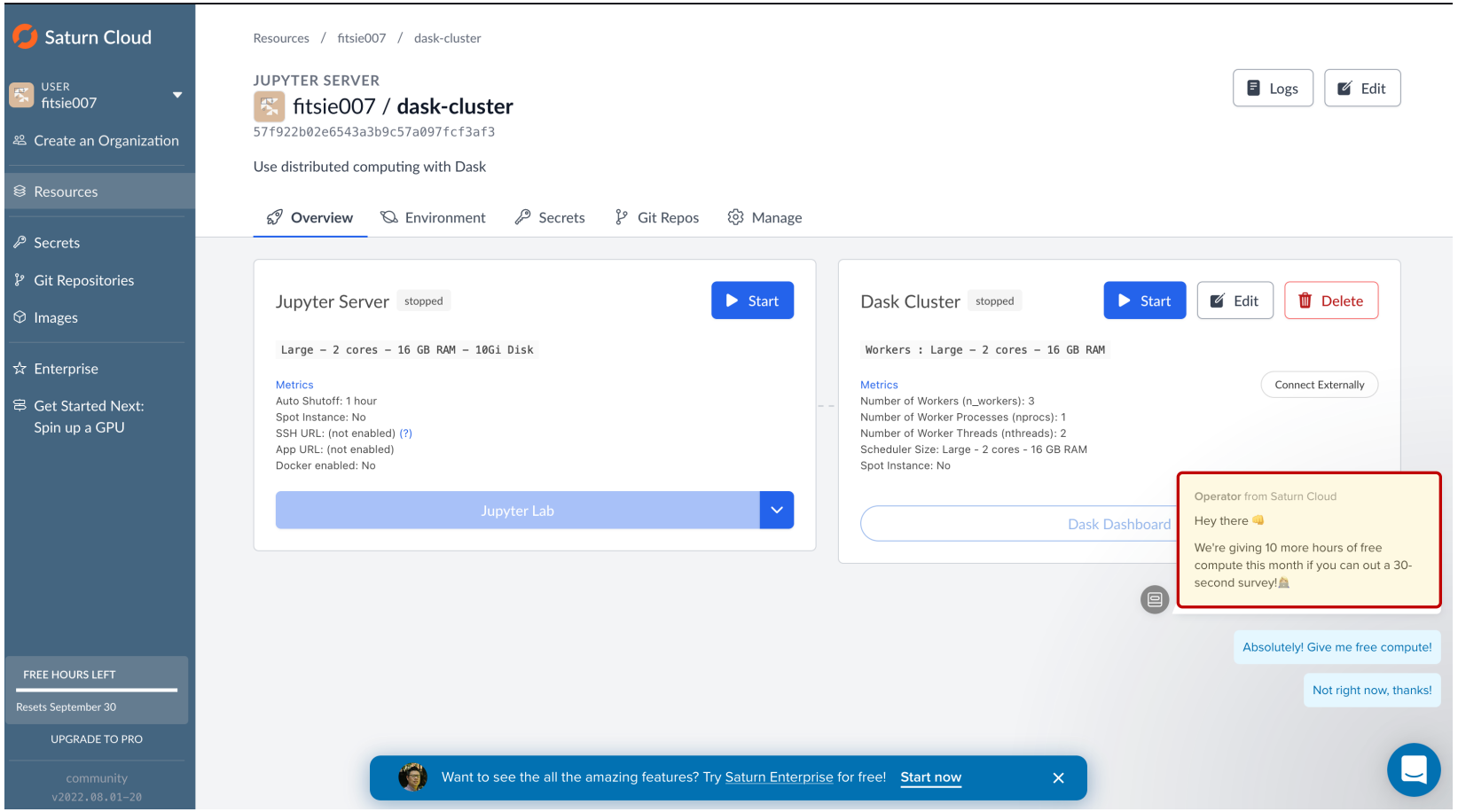
Create Resource

Enter a name for your resource such as dask-cluster.



Opt-in for Perks

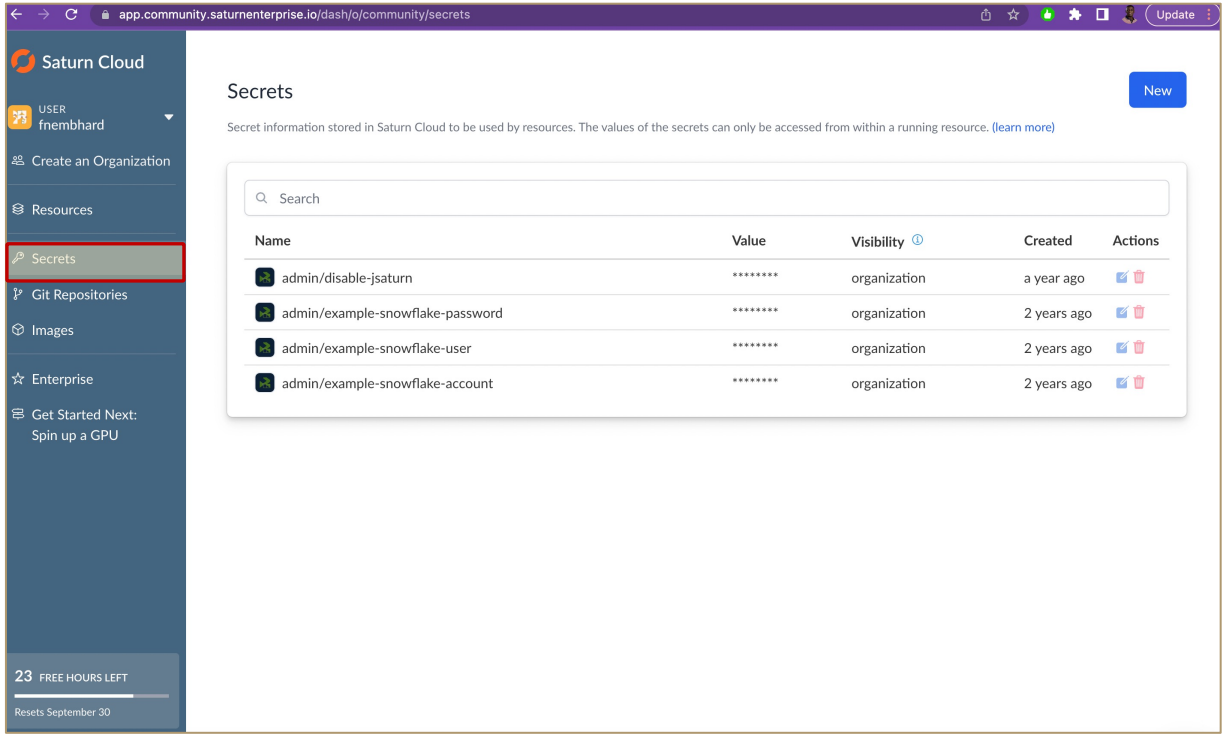
Look for ads that may give you a chance to get more time on the cluster.



FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Add AWS Credentials to Saturn

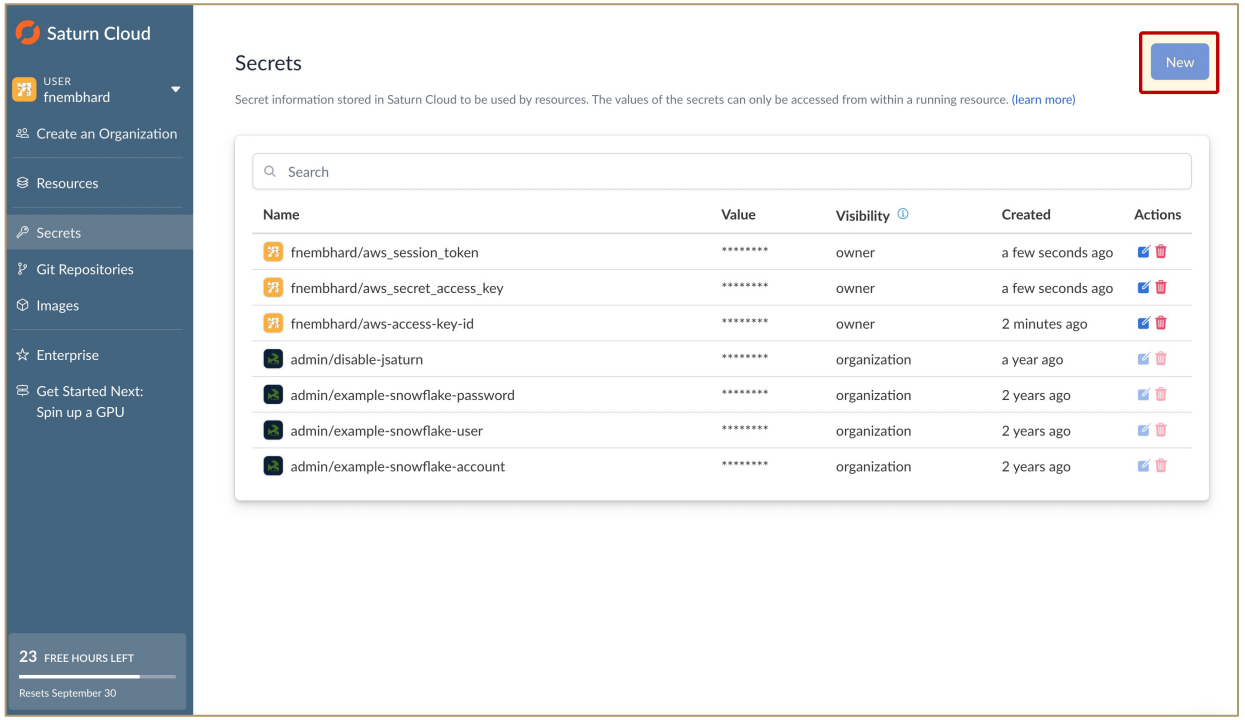
Go to your Saturn Cloud page (app.community.saturnenterprise.io) and click secrets



FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Add AWS Credentials to Saturn

Click New to add an AWS secret/variable.

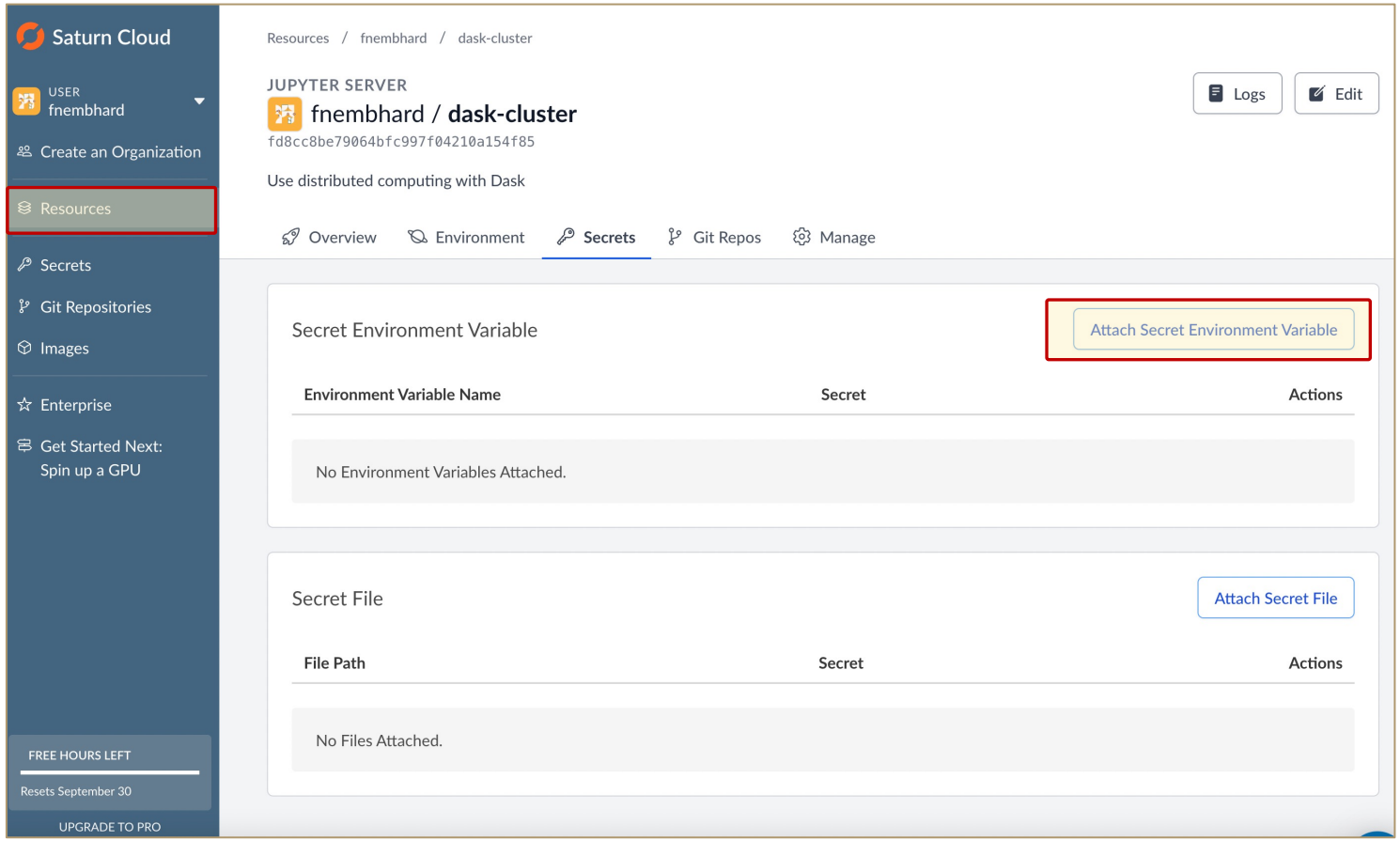


FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Attach secret to Resource

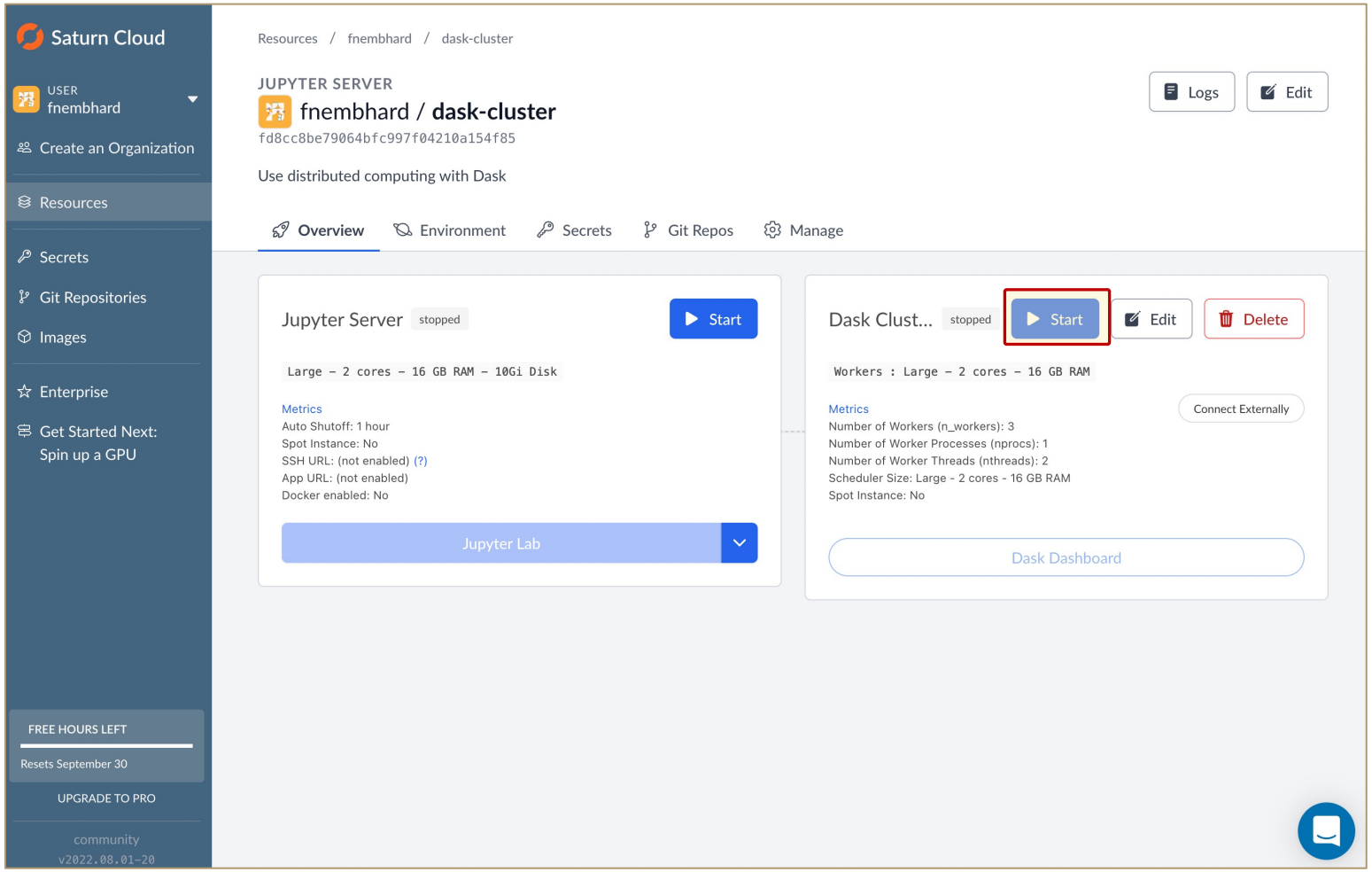
Click Resources and locate your dask cluster.

Click Attach Secret Environment Variable to attach the variable to the cluster environment.



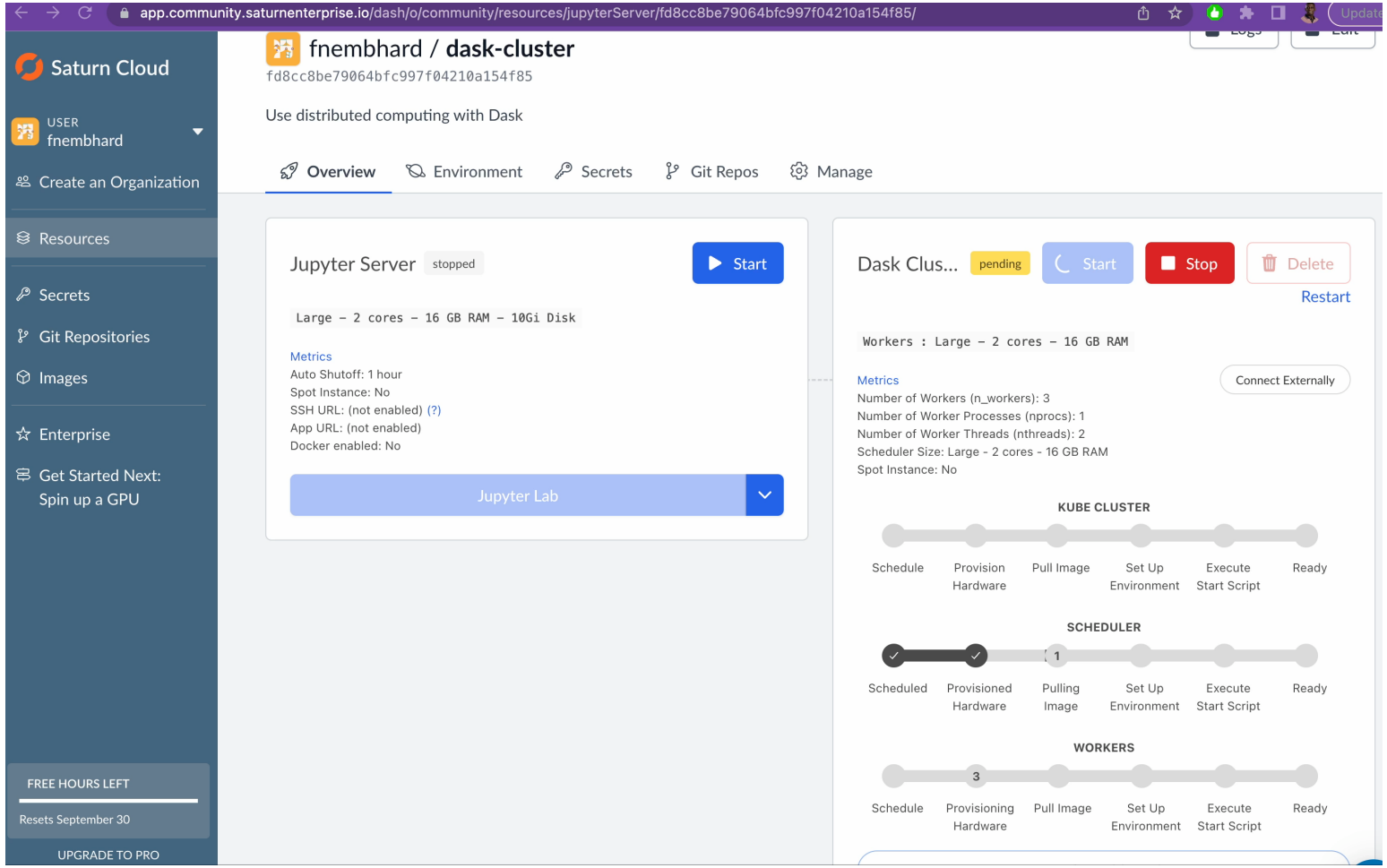
Start the cluster

Click Start to start the cluster



Start the cluster

The process may take several minutes to launch the dask scheduler and workers.



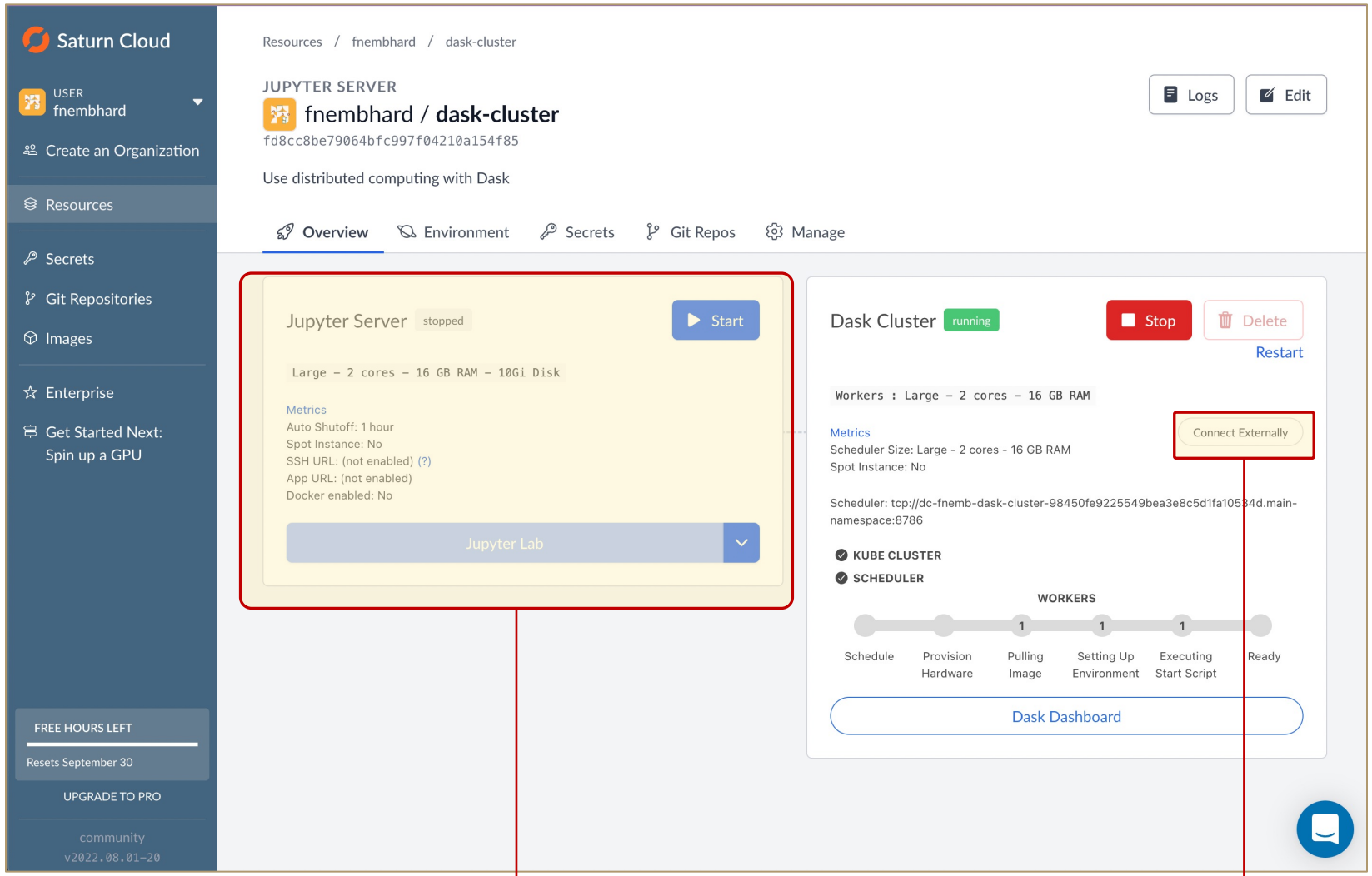
FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Access the cluster

You may access the cluster via Jupyter Server, which is available as a resource in Saturn Cloud or you may connect to the cluster remotely using your own environment.

To connect externally, you will need to build a conda environment.

Click Connect externally to access the cluster remotely.

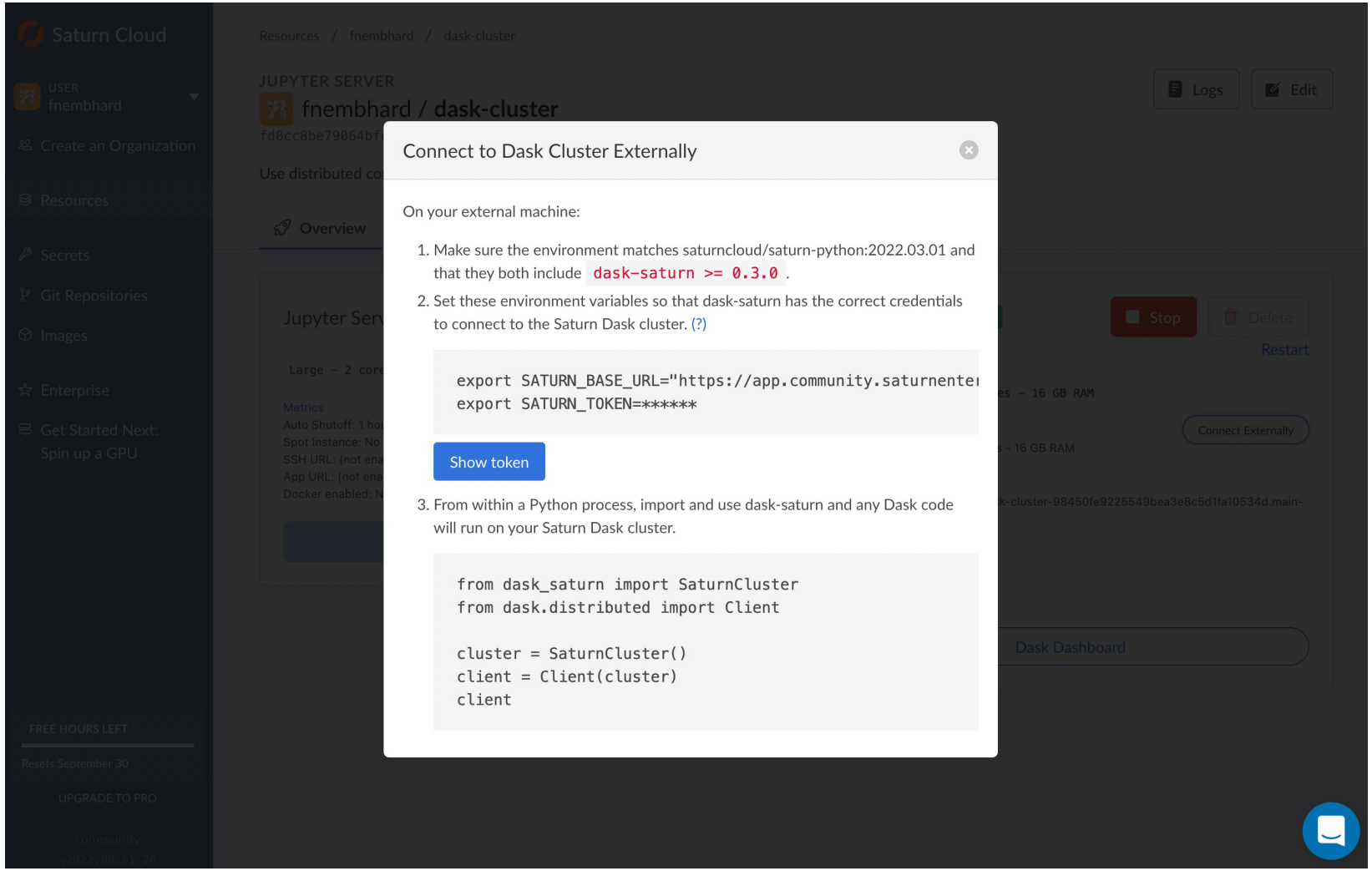


Choose one or the other

Connect to Cluster Remotely

Set up the given Saturn environment variables.

You may also set up the environment variables in your notebook. See the next 2 slides.



Create dask-saturn environment

To match the versions of the required packages, it is best to use the docker images provided or the environments.yml file to create the conda environment.

Locate the images here:
https://github.com/saturncloud/images/tree/release-2022.06.01

After downloading an image file, you can create the conda environment using any of the following commands:

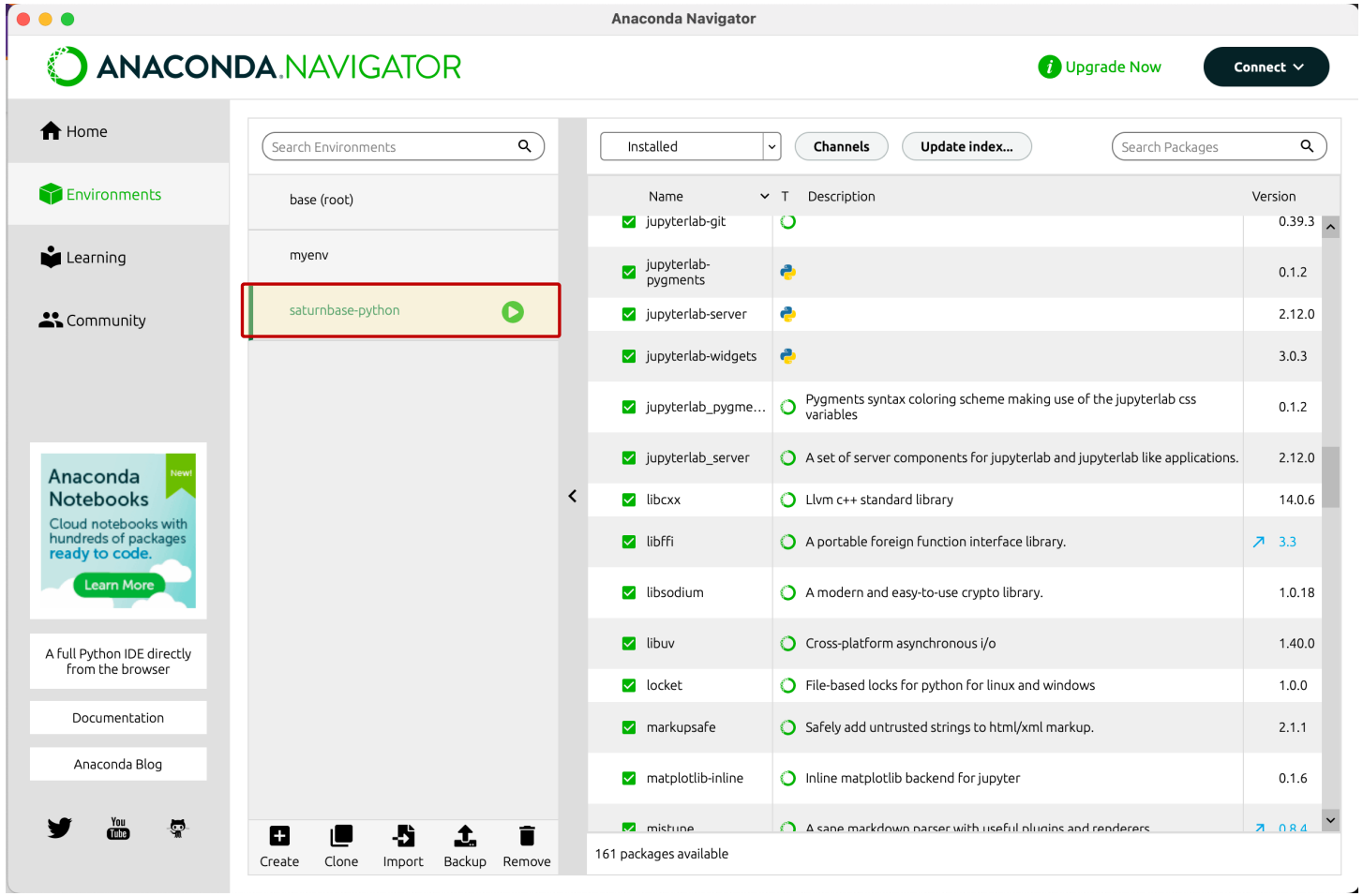
conda env create --name saturnbase-python --file=environment.yml

conda env create --file=environment.yml (leave off the environment name if it is present in the yml file)

Read more here:

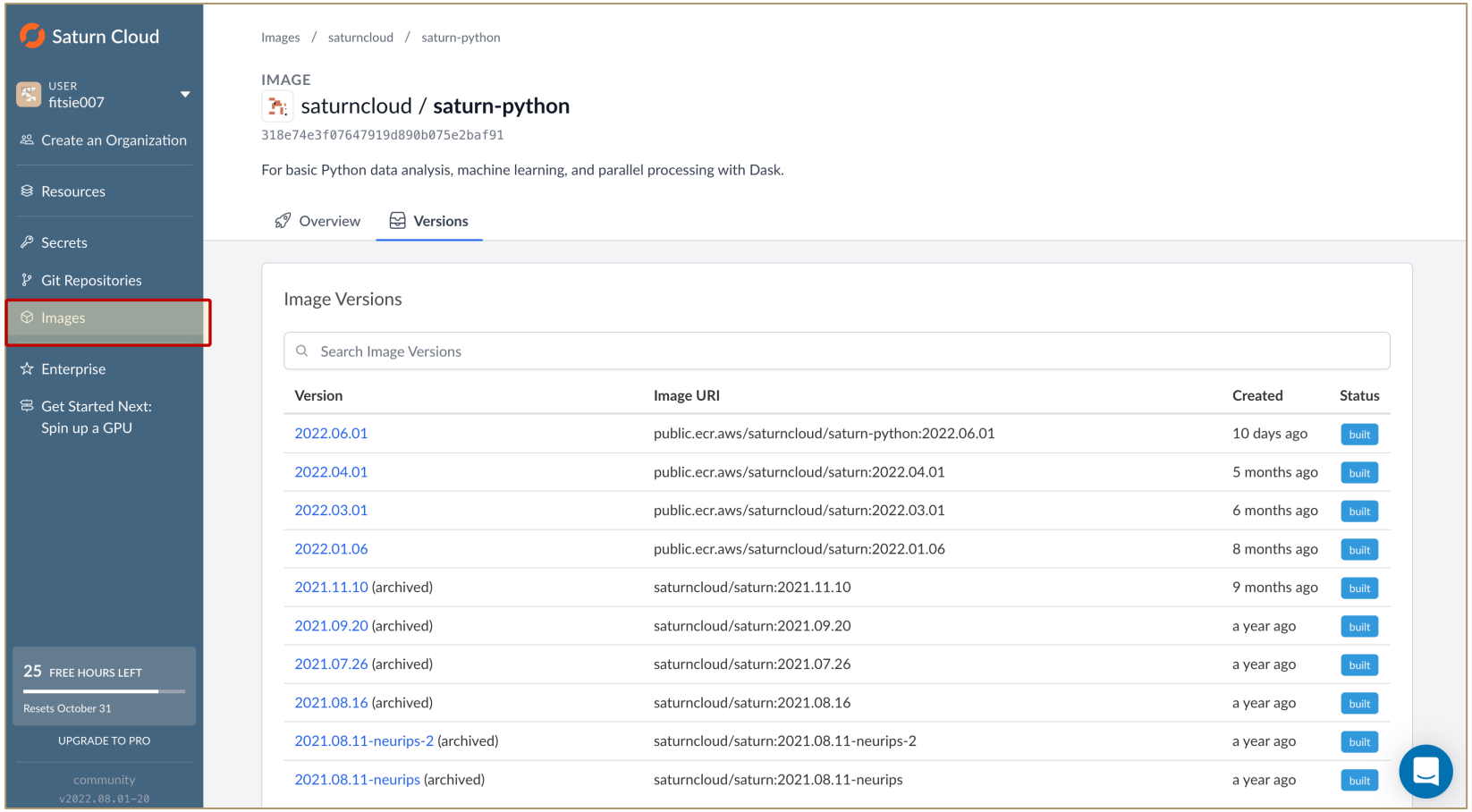
https://saturncloud.io/docs/using-saturn-cloud/external-connect/external_connect/

Important: Make sure the versions of the packages you install match the versions in the cloud.



Create dask-saturn environment

You can see the active images on Saturn Cloud.



FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Connect to and Use the Cluster.

For best results, create a conda environment with the following commands in order to run Saturn cloud successfully.

In [1]:
```
%env SATURN_BASE_URL=https://app.community.saturnenterprise.io
%env SATURN_TOKEN=server-IcantLetYouThrowYourselfAwayIcantLetYou
```

```
env: SATURN_BASE_URL=https://app.community.saturnenterprise.io
env: SATURN_TOKEN=server-IcantLetYouThrowYourselfAwayIcantLetYou
```

In [2]:
```
from dask_saturn import SaturnCluster
from dask.distributed import Client

cluster = SaturnCluster()
client = Client(cluster)
client
```

```
INFO:dask-saturn:Cluster is ready
INFO:dask-saturn:Registering default plugins
/Users/fitzroi/opt/anaconda3/lib/python3.8/site-packages/distributed/client.py:1105: VersionMismatchWarning: Mismatched versions found

+-------------+---------------+----------------+----------------+
| Package     | client        | scheduler      | workers        |
+-------------+---------------+----------------+----------------+
| cloudpickle | 1.6.0         | 2.0.0          | 2.0.0          |
| dask        | 2021.08.0     | 2021.07.2      | 2021.07.2      |
| distributed | 2021.08.0     | 2021.07.2      | 2021.07.2      |
| msgpack     | 1.0.2         | 1.0.3          | 1.0.3          |
| numpy       | 1.20.3        | 1.21.5         | 1.21.5         |
| pandas      | 1.3.1         | 1.4.1          | 1.4.1          |
| python      | 3.8.2.final.0 | 3.9.10.final.0 | 3.9.10.final.0 |
| toolz       | 0.11.1        | 0.11.2         | 0.11.2         |
+-------------+---------------+----------------+----------------+
Notes:
-  msgpack: Variation is ok, as long as everything is above 0.6
  warnings.warn(version_module.VersionMismatchWarning(msg[0]["warning"]))
INFO:dask-saturn:Success!
```

Out[2]:

Client

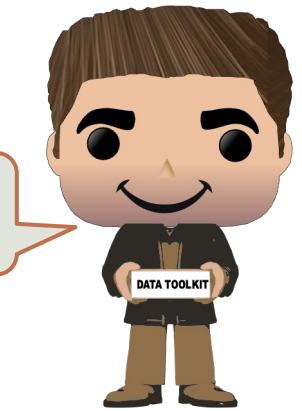
Client-2ab664f2-3e9f-11ed-9e99-92d1bdd47554

Connection method: Cluster object

Cluster type: SaturnCluster

Dashboard: https://dc-fnemb-dask-cluster-98450fe9225549bea3e8c5d1fa10534d.community.saturnenterprise.io (https://dc-fnemb-dask-cluster-98450fe9225549bea3e8c5d1fa10534d.community.saturnenterprise.io)

Cluster Info

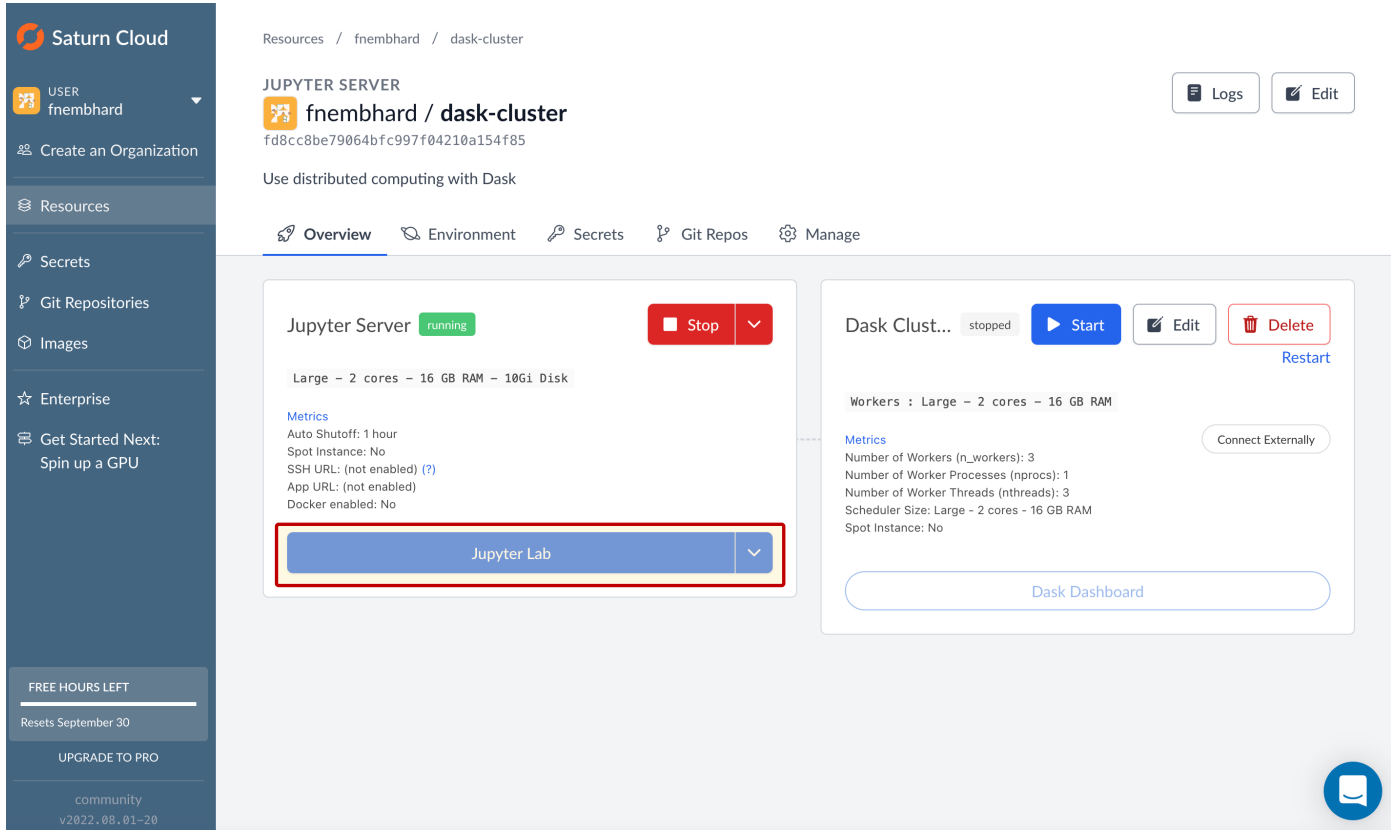


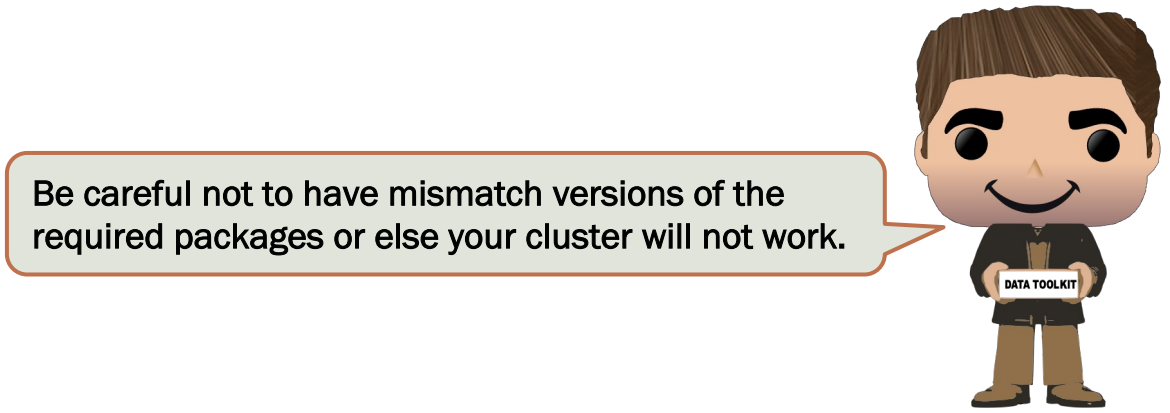
Be careful not to have mismatch versions of the required packages or else your cluster will not work.

FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Connecting to the cluster from Jupyter via Saturn Cloud.

Start Jupyter Lab and execute your python code via the notebook.

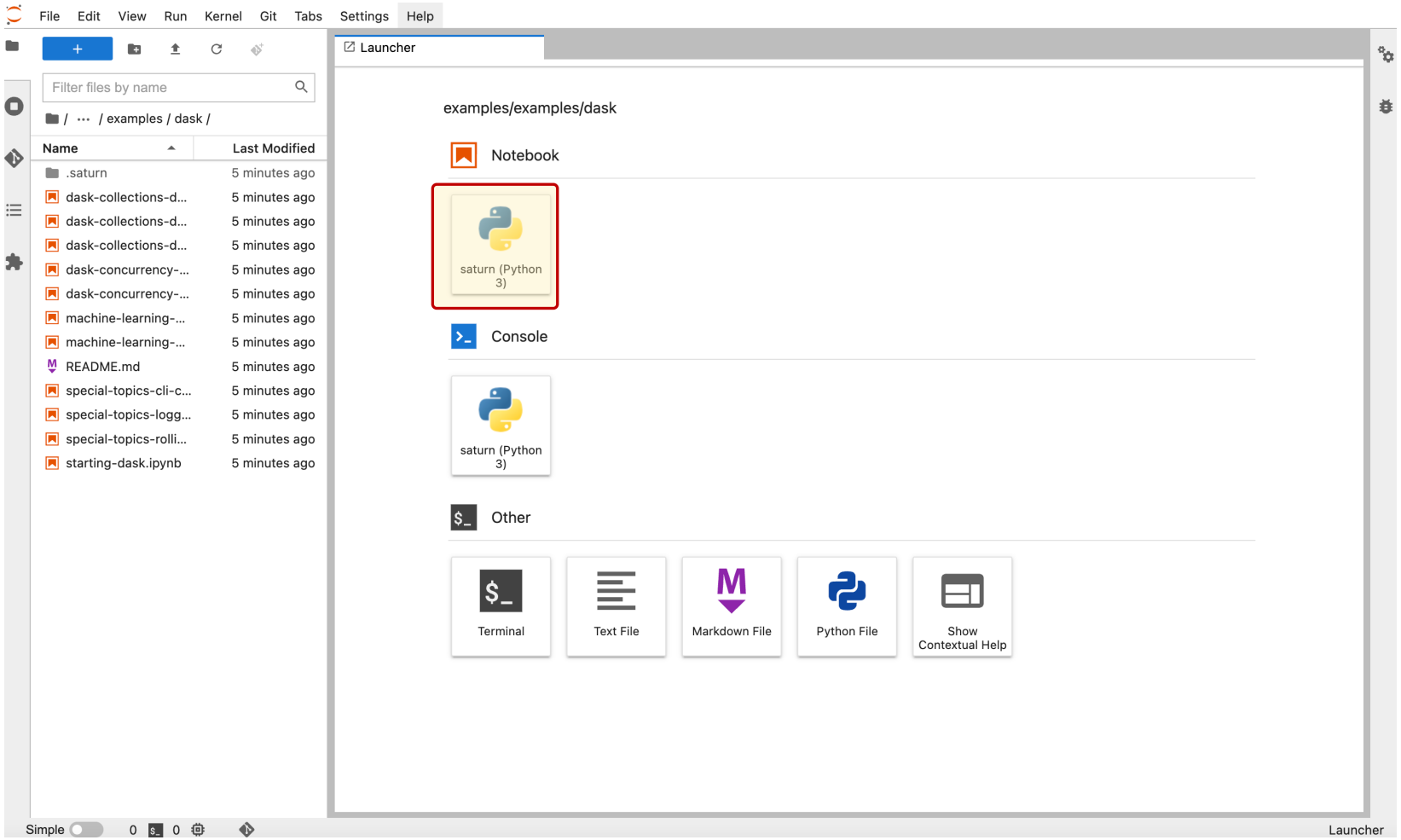




FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

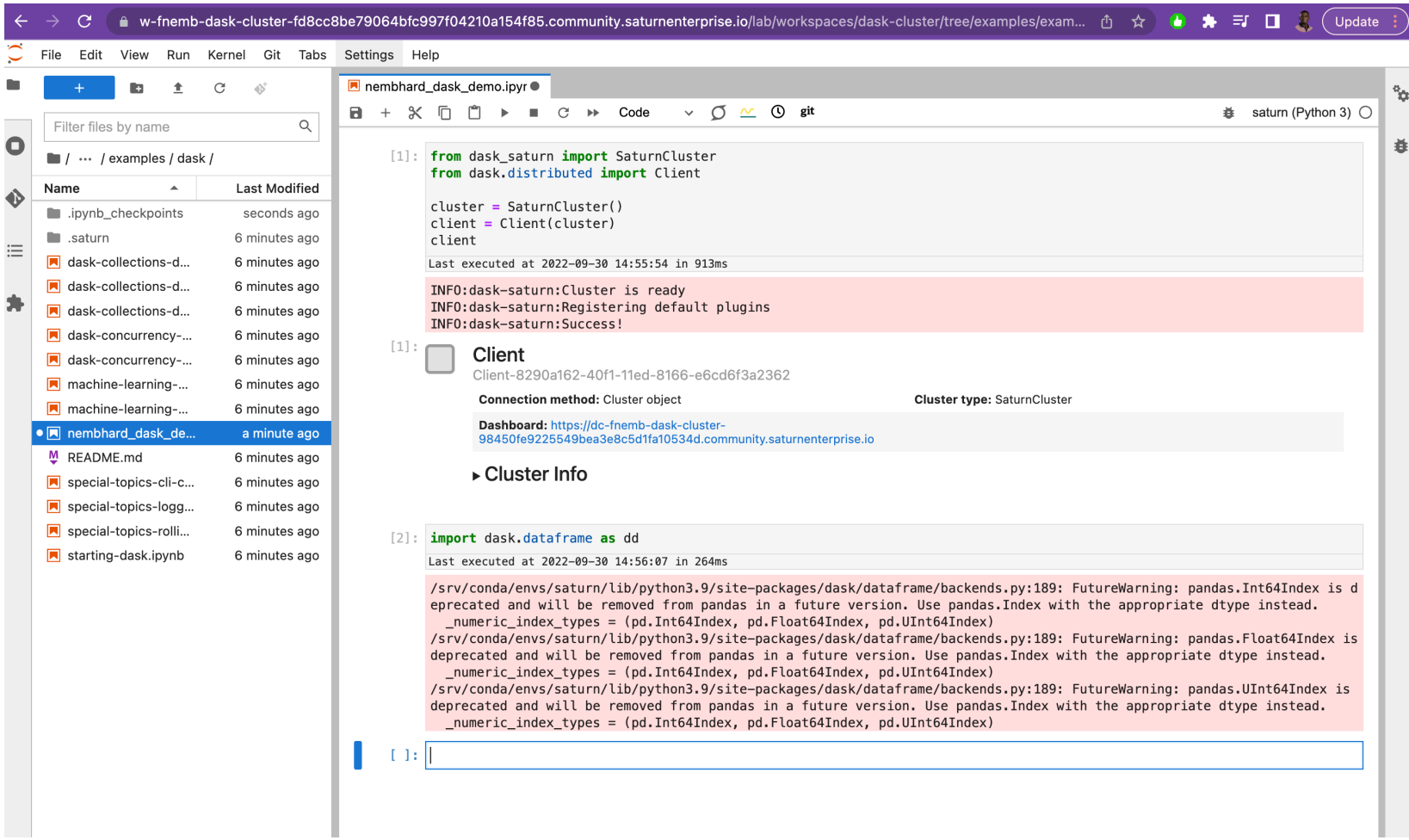
Connect to and Use the Cluster.

One disadvantage of using Jupyter Lab from Saturn Cloud is that it will eat into your allotted time. It may be best to connect remotely to the cluster using your own conda environment.



Connect to and Use the Cluster.

This screenshot demonstrates how to connect to the cluster and code from Jupyter Lab in Saturn Cloud. Some warnings may appear.



FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

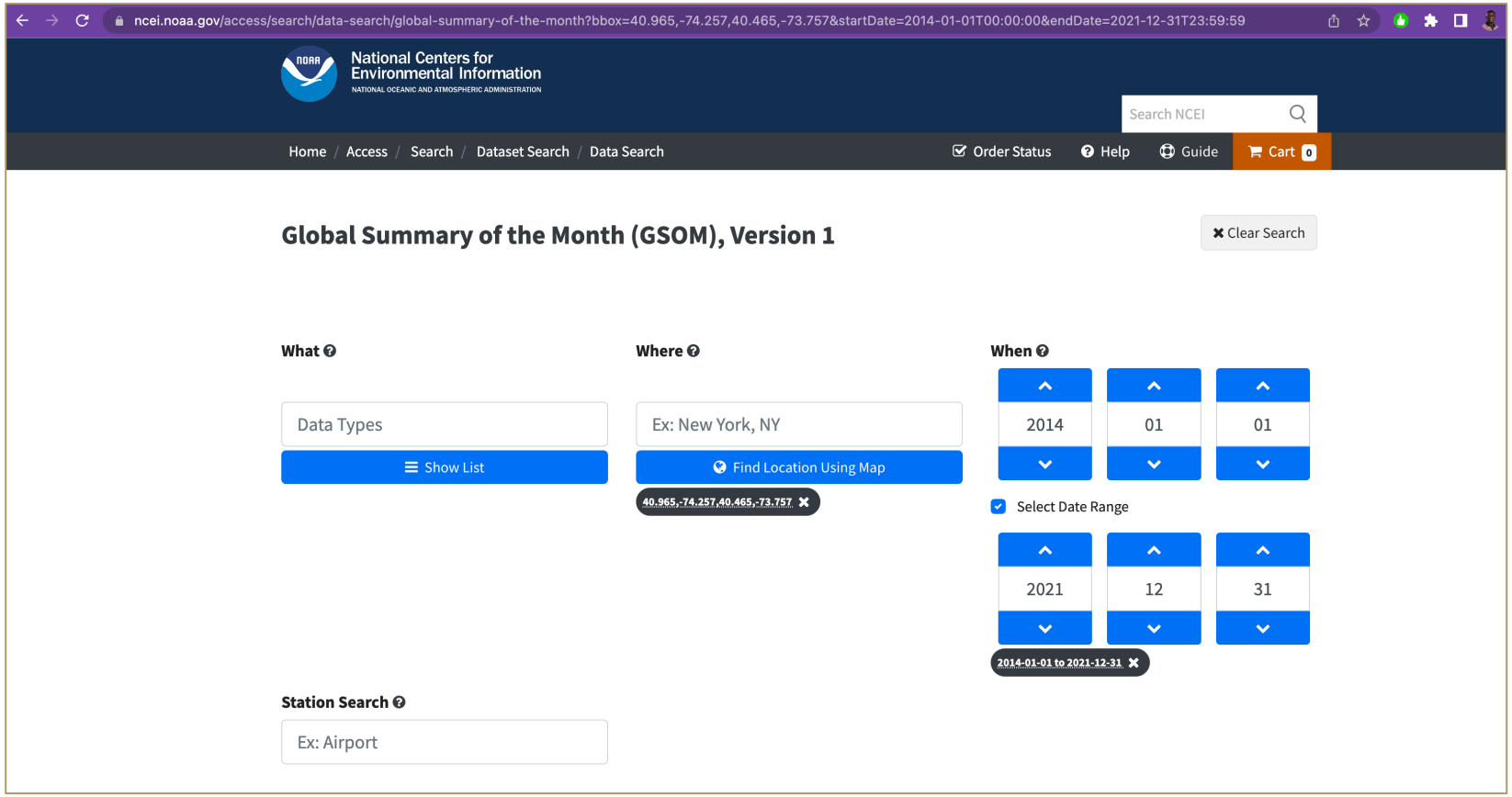
Bonus: Getting Temperature Data from NOAA

THIS SECTION CAPTURES BRIEFLY HOW TO ACCCESS NOAA'S API TO RETRIEVE MONTHLY TEMPERATURE DATA.

Augmenting Ticket Data with Weather Data

The following page allows one to download GSOM datasets from NOAA.

https://www.ncei.noaa.gov/access/search/data-search/global-summary-of-the-month

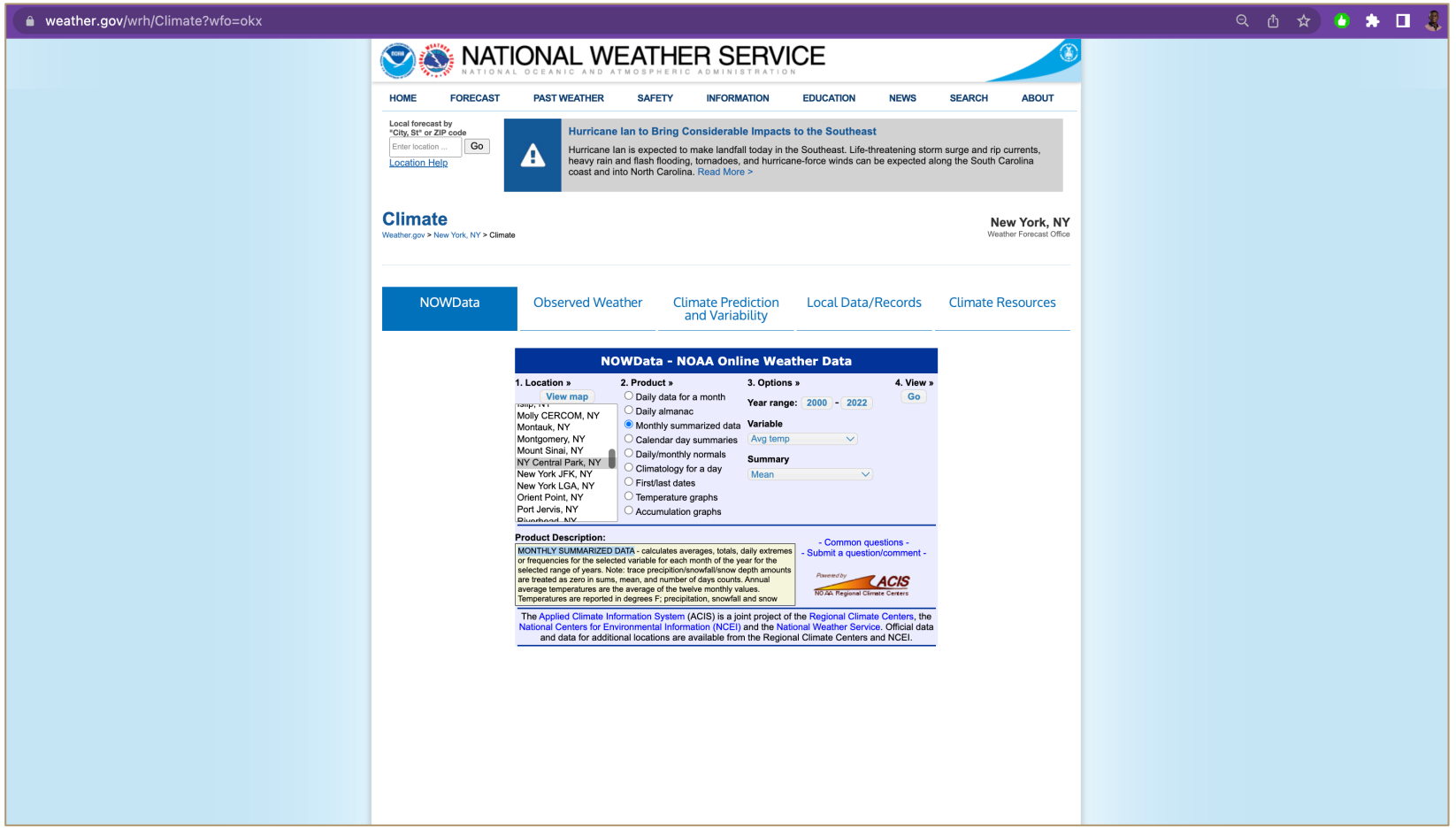


FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Augmenting Ticket Data with Weather Data

The following page allows one to view a table of data such as temperatures over a period of time.

https://www.weather.gov/wrh/Climate?wfo=okx

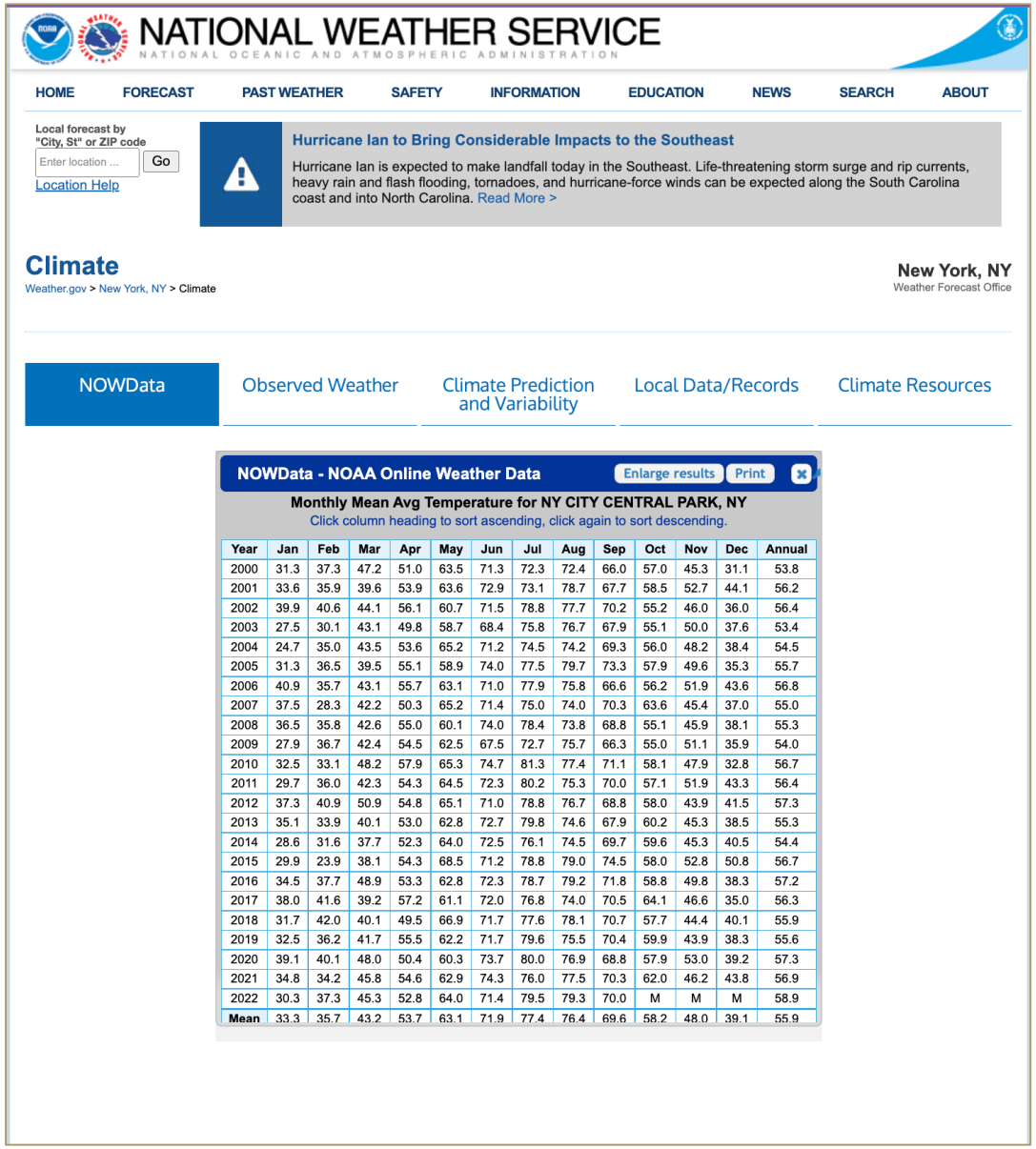


FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Augmenting Ticket Data with Weather Data

The following page allows one to view a table of data such as temperatures over a period of time.

https://www.weather.gov/wrh/Climate?wfo=okx



NATIONAL WEATHER SERVICE
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION

HOME | FORECAST | PAST WEATHER | SAFETY | INFORMATION | EDUCATION | NEWS | SEARCH | ABOUT

Local forecast by "City, St" or ZIP code
Enter location ... Go
Location Help

Hurricane Ian to Bring Considerable Impacts to the Southeast

Hurricane Ian is expected to make landfall today in the Southeast. Life-threatening storm surge and rip currents, heavy rain and flash flooding, tornadoes, and hurricane-force winds can be expected along the South Carolina coast and into North Carolina. Read More >

Climate

Weather.gov > New York, NY > Climate

New York, NY
Weather Forecast Office

NOWData | Observed Weather | Climate Prediction and Variability | Local Data/Records | Climate Resources

NOWData - NOAA Online Weather Data Enlarge results Print

Monthly Mean Avg Temperature for NY CITY CENTRAL PARK, NY

Click column heading to sort ascending, click again to sort descending.

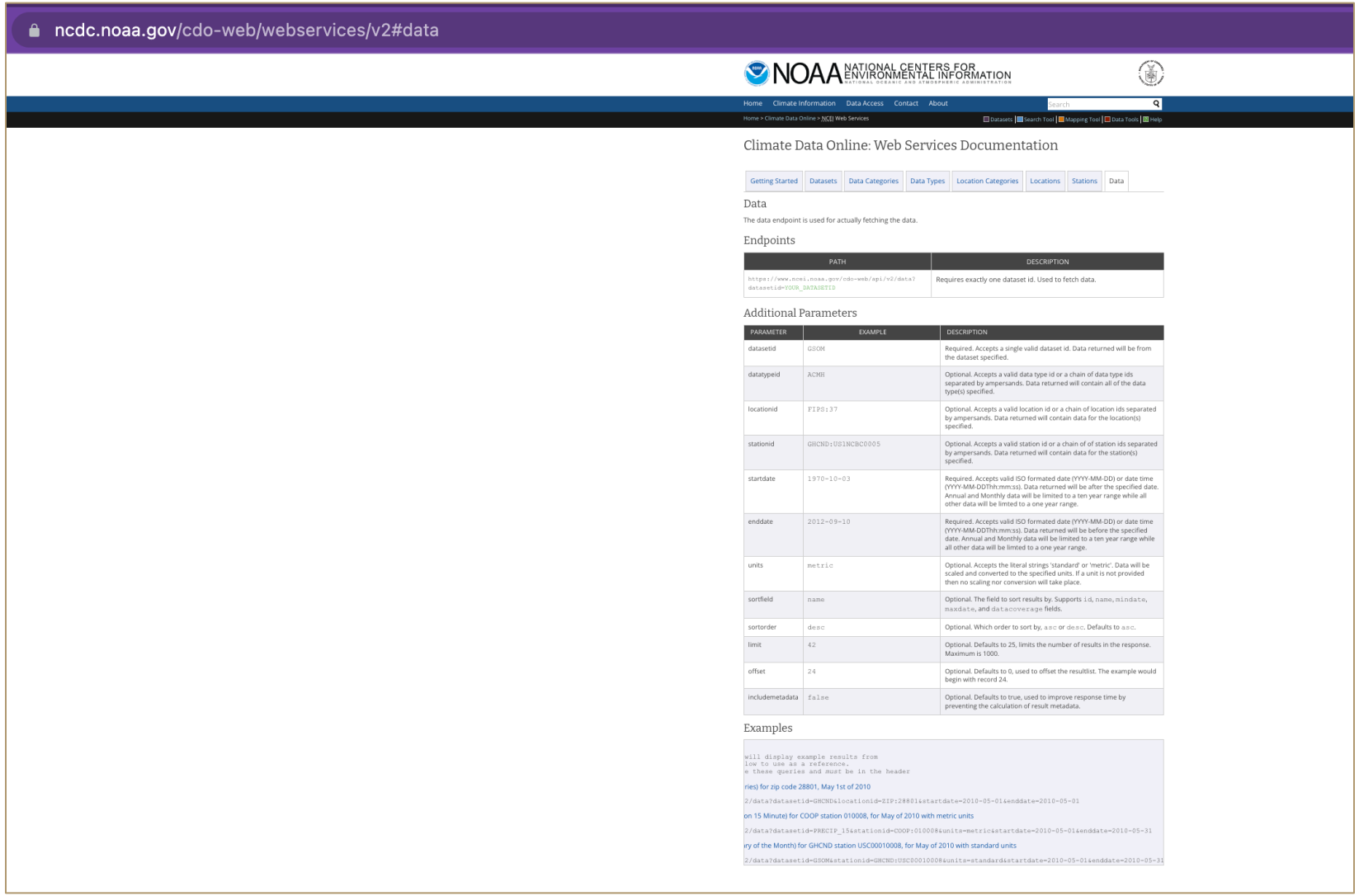
Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Annual
2000	31.3	37.3	47.2	51.0	63.5	71.3	72.3	72.4	66.0	57.0	45.3	31.1	53.8
2001	33.6	35.9	39.6	53.9	63.6	72.9	73.1	78.7	67.7	58.5	52.7	44.1	56.2
2002	39.9	40.6	44.1	56.1	60.7	71.5	78.8	77.7	70.2	55.2	46.0	36.0	56.4
2003	27.5	30.1	43.1	49.8	58.7	68.4	75.8	76.7	67.9	55.1	50.0	37.6	53.4
2004	24.7	35.0	43.5	53.6	65.2	71.2	74.5	74.2	69.3	56.0	48.2	38.4	54.5
2005	31.3	36.5	39.5	55.1	58.9	74.0	77.5	79.7	73.3	57.9	49.6	35.3	55.7
2006	40.9	35.7	43.1	55.7	63.1	71.0	77.9	75.8	66.6	56.2	51.9	43.6	56.8
2007	37.5	28.3	42.2	50.3	65.2	71.4	75.0	74.0	70.3	63.6	45.4	37.0	55.0
2008	36.5	35.8	42.6	55.0	60.1	74.0	78.4	73.8	68.8	55.1	45.9	38.1	55.3
2009	27.9	36.7	42.4	54.5	62.5	67.5	72.7	75.7	66.3	55.0	51.1	35.9	54.0
2010	32.5	33.1	48.2	57.9	65.3	74.7	81.3	77.4	71.1	58.1	47.9	32.8	56.7
2011	29.7	36.0	42.3	54.3	64.5	72.3	80.2	75.3	70.0	57.1	51.9	43.3	56.4
2012	37.3	40.9	50.9	54.8	65.1	71.0	78.8	76.7	68.8	58.0	43.9	41.5	57.3
2013	35.1	33.9	40.1	53.0	62.8	72.7	79.8	74.6	67.9	60.2	45.3	38.5	55.3
2014	28.6	31.6	37.7	52.3	64.0	72.5	76.1	74.5	69.7	59.6	45.3	40.5	54.4
2015	29.9	23.9	38.1	54.3	68.5	71.2	78.8	79.0	74.5	58.0	52.8	50.8	56.7
2016	34.5	37.7	48.9	53.3	62.8	72.3	78.7	79.2	71.8	58.8	49.8	38.3	57.2
2017	38.0	41.6	39.2	57.2	61.1	72.0	76.8	74.0	70.5	64.1	46.6	35.0	56.3
2018	31.7	42.0	40.1	49.5	66.9	71.7	77.6	78.1	70.7	57.7	44.4	40.1	55.9
2019	32.5	36.2	41.7	55.5	62.2	71.7	79.6	75.5	70.4	59.9	43.9	38.3	55.6
2020	39.1	40.1	48.0	50.4	60.3	73.7	80.0	76.9	68.8	57.9	53.0	39.2	57.3
2021	34.8	34.2	45.8	54.6	62.9	74.3	76.0	77.5	70.3	62.0	46.2	43.8	56.9
2022	30.3	37.3	45.3	52.8	64.0	71.4	79.5	79.3	70.0	M	M	M	58.9
Mean	33.3	35.7	43.2	53.7	63.1	71.9	77.4	76.4	69.6	58.2	48.0	39.1	55.9

FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Augmenting Ticket Data with Weather Data

Use the following NOAA API to access temperature data:

https://www.ncdc.noaa.gov/cdo-web/webservices/v2#data



References

The following references were used to create this tutorial.

- AWS Academy
- https://boto3.amazonaws.com/v1/documentation/api/latest/guide/s3-examples.html
- https://saturncloud.io/docs/using-saturn-cloud/external-connect/external_connect/
- https://saturncloud.io/docs/using-saturn-cloud/secrets/

FLORIDA TECH
FLORIDA'S STEM UNIVERSITY