

# CSE 4510/5310

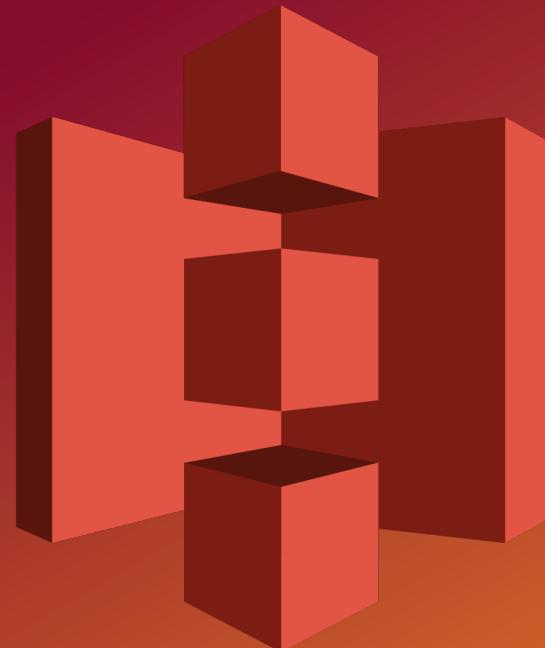
## Big Data

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**



# 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](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.

The screenshot shows the 'Saturn Cloud Plans and Pricing' page. At the top, there's a navigation bar with links for 'Why Saturn Cloud', 'Partners', 'Resources', 'Plans & Pricing', 'Login', and a prominent 'Start For Free' button. Below the navigation, the title 'Saturn Cloud Plans and Pricing' is displayed, followed by the subtitle 'Flexible plans for every data scientist, team, and company.' The page is divided into three main sections: 'Hosted', 'Hosted Organizations', and 'Enterprise'. Each section contains a brief description and a bulleted list of features, along with 'Start for free' and 'Learn more' buttons. The 'Hosted' section includes a note about free RAM and GPU instances, while the 'Enterprise' section highlights advanced security and custom VPCs.

Hosted	Hosted Organizations	Enterprise
<p>The essentials for individual data scientists</p> <ul style="list-style-type: none"><li>• 30 hours a month free of 64GB RAM and GPU instances</li><li>• Use JupyterLab, RStudio, and Dask</li><li>• Upgrade to a paid account for:<ul style="list-style-type: none"><li>◦ Deployments and jobs</li><li>◦ Up to 4TB of RAM and 8 GPUs</li><li>◦ \$20 a month in free credits</li></ul></li></ul>	<p>Collaboration tools on our cloud environment</p> <ul style="list-style-type: none"><li>• Share resources between users on a team</li><li>• Create group-owned resources</li><li>• Use admin tools to monitor and control usage</li></ul>	<p>Install Saturn Cloud into your own AWS account</p> <ul style="list-style-type: none"><li>• Installs into your AWS account from the AWS Marketplace</li><li>• Advanced security: SSO and installation into custom VPCs and private subnets available</li><li>• Dedicated technical support</li></ul>

# Create an Account

Choose your favorite option to create an account.

 SaturnCloud

**Create your Saturn Cloud account**  
Start for free in seconds—No credit card required


---

OR

**Username**

**Email**

**Sign Up**

Already have an account? [Login](#)

By using this service, you agree to our [Privacy Policy](#) and [Terms of Service](#)

# Sign In

After logging in, you should see the following screen.

Click **New Resource from a Template**

The screenshot shows the Saturn Cloud web interface at [app.community.saturnenterprise.io/dash/o/community/resources](https://app.community.saturnenterprise.io/dash/o/community/resources). The left sidebar includes options like 'Create an Organization', 'Resources' (which is selected), 'Secrets', 'Git Repositories', 'Images', 'Enterprise', and 'Get Started Next: Make a Workspace'. A 'FREE HOURS LEFT' banner indicates resets on September 30. The main content area has a heading 'Create a Resource' and four cards: 'New Python Server' (using JupyterLab or PyCharm), 'New RStudio Server' (writing/run code in RStudio IDE), 'New Deployment' (hosting an app/API), and 'New Job' (running tasks on schedule). Below these is a 'New Resource from a Template' section with a 'TensorFlow', 'snowflake', 'dask', and 'PyTorch' logo row. The bottom section shows a 'Resources' list with a search bar, sorting options ('Show All Resource Types', 'Sort By', 'Recently Started'), and a 'Create a new resource' button. A Florida Tech watermark is visible in the bottom right.

# Select Dask

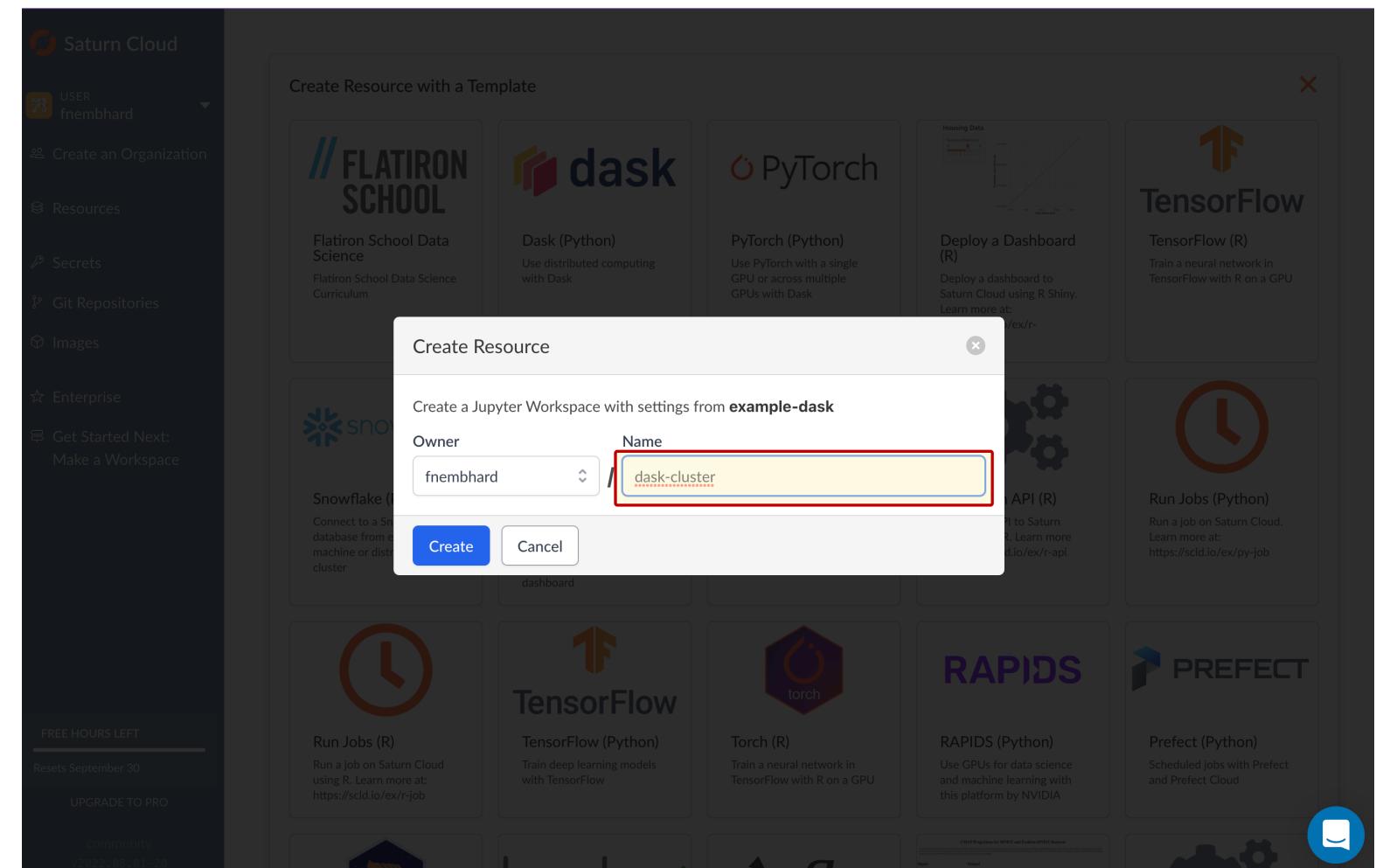
Click **dask** to access a resource template

The screenshot shows the Saturn Cloud web interface. On the left is a sidebar with user information (fnembhard), navigation links (Create an Organization, Resources, Secrets, Git Repositories, Images, Enterprise), and a 'Get Started Next: Make a Workspace' button. At the bottom of the sidebar is a 'FREE HOURS LEFT' bar and an 'UPGRADE TO PRO' button. The main area is titled 'Create Resource with a Template' and displays a grid of 12 resource templates. The 'dask' template is highlighted with a red border. Other templates include PyTorch, TensorFlow, Snowflake, FastAPI, Run Jobs (Python/R), TensorFlow (Python), Torch (R), RAPIDS (Python), and Prefect (Python). Each template has a title, icon, and a brief description.

Template	Description
<b>dask</b>	Dask (Python) Use distributed computing with Dask
<b>PyTorch</b>	PyTorch (Python) Use PyTorch with a single GPU or across multiple GPUs with Dask
<b>TensorFlow</b>	TensorFlow (R) Train a neural network in TensorFlow with R on a GPU
<b>Snowflake</b>	Snowflake (Python) Connect to a Snowflake database from either a single machine or distributed cluster
<b>FastAPI</b>	Deploy an API (Python) Deploy an API to Saturn Cloud. Learn more at: <a href="https://scld.io/ex/py-api">https://scld.io/ex/py-api</a>
<b>Run Jobs (Python)</b>	Deploy a Dashboard (Python) Deploy a dashboard to Saturn Cloud. Learn more at: <a href="https://scld.io/ex/py-dashboard">https://scld.io/ex/py-dashboard</a>
<b>TensorFlow</b>	TensorFlow (Python) Train deep learning models with TensorFlow
<b>Torch (R)</b>	Torch (R) Train a neural network in TensorFlow with R on a GPU
<b>RAPIDS</b>	RAPIDS (Python) Use GPUs for data science and machine learning with this platform by NVIDIA
<b>PREFECT</b>	Prefect (Python) Scheduled jobs with Prefect and Prefect Cloud
<b>Run Jobs (R)</b>	Run Jobs (R) Run a job on Saturn Cloud using R. Learn more at: <a href="https://scld.io/ex/r-job">https://scld.io/ex/r-job</a>

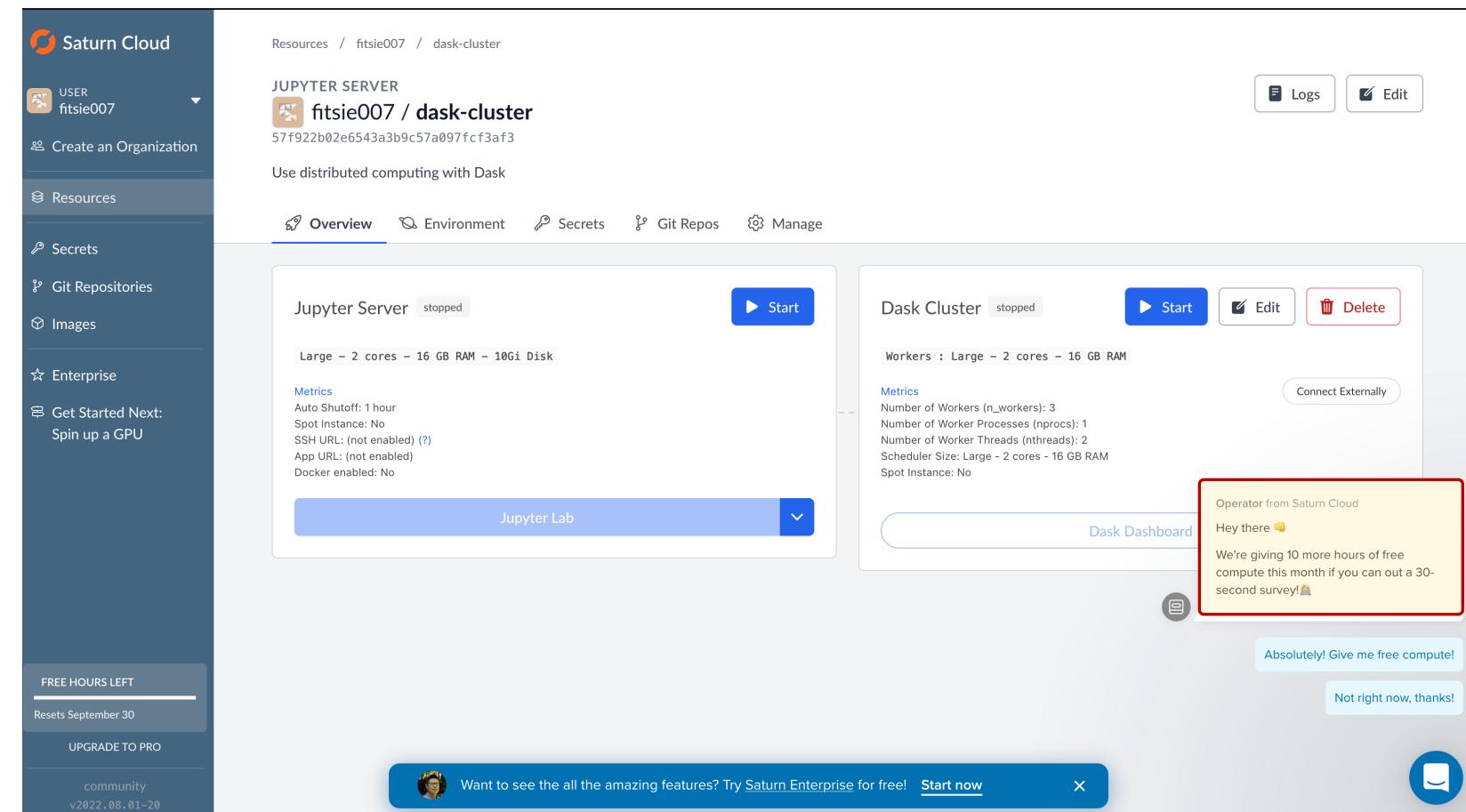
# Create Resource

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.



The screenshot shows the Saturn Cloud web interface. On the left is a sidebar with user information (USER fitsie007), navigation links (Create an Organization, Resources, Secrets, Git Repositories, Images, Enterprise), and a call-to-action (Get Started Next: Spin up a GPU). A "FREE HOURS LEFT" section indicates Resets September 30. Below that is an "UPGRADE TO PRO" button and version information (community v2022.08.01-20).

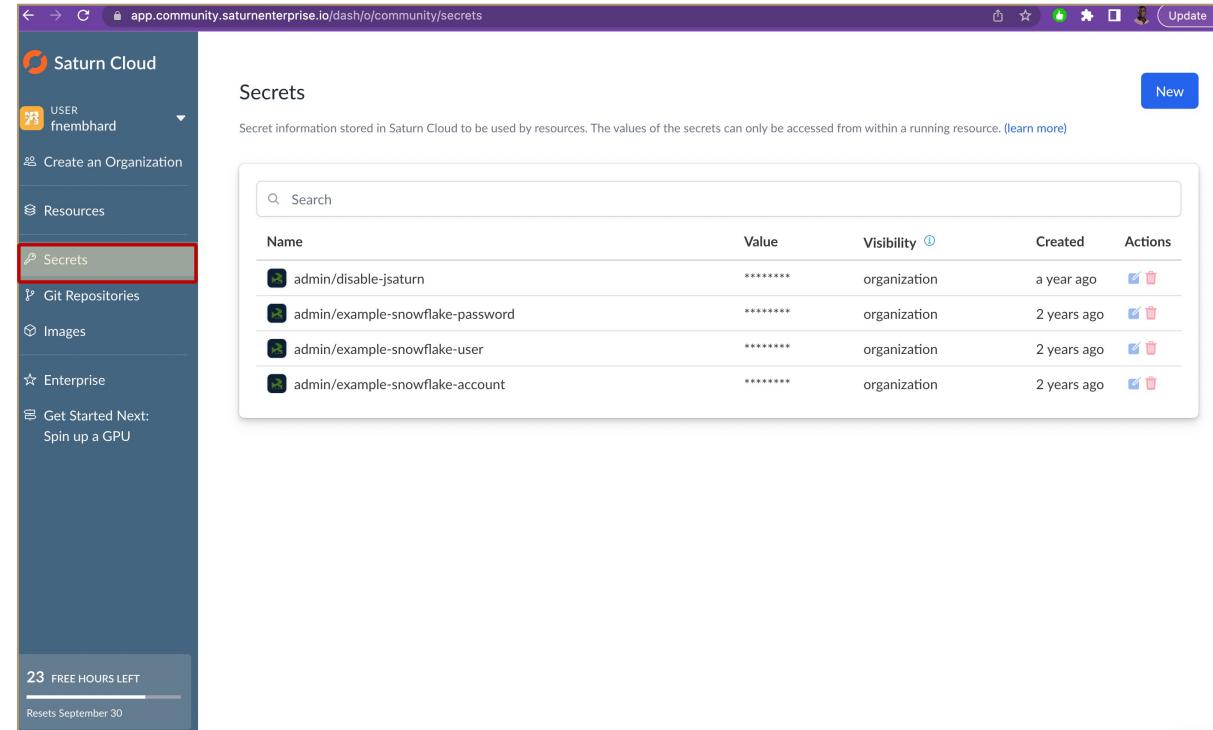
The main content area is titled "Resources / fitsie007 / dask-cluster". It displays two resources:

- JUPYTER SERVER**: fitsie007 / **dask-cluster** (57f922b02e6543a3b9c57a097fcf3af3). Status: stopped. Metrics: Auto Shutoff: 1 hour, Spot Instance: No, SSH URL: (not enabled) (?), App URL: (not enabled), Docker enabled: No. A "Jupyter Lab" button is present.
- Dask Cluster**: status: stopped. Metrics: Workers : Large - 2 cores - 16 GB RAM. Number of Workers (n\_workers): 3, Number of Worker Processes (nprocs): 1, Number of Worker Threads (nthreads): 2, Scheduler Size: Large - 2 cores - 16 GB RAM, Spot Instance: No. Buttons: Start, Edit, Delete. A "Dask Dashboard" button is present.

A red box highlights a message from an "Operator from Saturn Cloud": "Hey there! We're giving 10 more hours of free compute this month if you can out a 30-second survey! 🎉". Below it are buttons for "Absolutely! Give me free compute!" and "Not right now, thanks!". A blue speech bubble icon is in the bottom right corner.

# Add AWS Credentials to Saturn

Go to your Saturn Cloud page ([app.community.saturnenterprise.io](https://app.community.saturnenterprise.io)) and click secrets



The screenshot shows the Saturn Cloud web interface. On the left, a sidebar menu includes options like 'USER fnembhard', 'Create an Organization', 'Resources', 'Secrets' (which is highlighted with a red box), 'Git Repositories', 'Images', and 'Enterprise'. Below the sidebar, a message says 'Get Started Next: Spin up a GPU'. At the bottom, a banner indicates '23 FREE HOURS LEFT' and 'Resets September 30'. The main content area is titled 'Secrets' and contains a table of stored secrets:

Name	Value	Visibility	Created	Actions
admin/disable-jsaturn	*****	organization	a year ago	
admin/example-snowflake-password	*****	organization	2 years ago	
admin/example-snowflake-user	*****	organization	2 years ago	
admin/example-snowflake-account	*****	organization	2 years ago	

# Add AWS Credentials to Saturn

Click **New** to add an AWS secret/variable.

The screenshot shows the Saturn Cloud interface with a sidebar on the left and a main content area on the right.

**Sidebar:**

- USER fnembhard
- Create an Organization
- Resources
- Secrets** (highlighted)
- Git Repositories
- Images
- Enterprise
- Get Started Next: Spin up a GPU

**Main Content Area:**

### Secrets

Secret information stored in Saturn Cloud to be used by resources. The values of the secrets can only be accessed from within a running resource. [\(learn more\)](#)

Name	Value	Visibility ⓘ	Created	Actions
fnembhard/aws_session_token	*****	owner	a few seconds ago	
fnembhard/aws_secret_access_key	*****	owner	a few seconds ago	
fnembhard/aws-access-key-id	*****	owner	2 minutes ago	
admin/disable-jssaturn	*****	organization	a year ago	
admin/example-snowflake-password	*****	organization	2 years ago	
admin/example-snowflake-user	*****	organization	2 years ago	
admin/example-snowflake-account	*****	organization	2 years ago	

A red box highlights the **New** button in the top right corner of the main content area.

# Attach secret to Resource

Click **Resources** and locate your dask cluster.

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

The screenshot shows the Saturn Cloud web interface. On the left, a sidebar menu has 'Resources' highlighted with a red box. The main content area shows a 'JUPYTER SERVER' named 'fnembhard / dask-cluster'. Below the server name is a unique identifier: 'fd8cc8be79064bfc997f04210a154f85'. A descriptive text states 'Use distributed computing with Dask'. Below this are navigation tabs: Overview, Environment, Secrets (which is underlined in blue), Git Repos, and Manage. The 'Secrets' section contains two tables: 'Secret Environment Variable' and 'Secret File'. Both tables have a header row with columns: Environment Variable Name, Secret, and Actions. The first table's body says 'No Environment Variables Attached.' and has a 'Attach Secret Environment Variable' button in its top right corner, also highlighted with a red box. The second table's body says 'No Files Attached.' and has a 'Attach Secret File' button in its top right corner.

# Start the cluster

Click **Start** to start the cluster

The screenshot shows the Saturn Cloud interface. On the left is a sidebar with the user's name (fnembhard), a 'Create an Organization' button, 'Resources', 'Secrets', 'Git Repositories', 'Images', 'Enterprise' (with 'Get Started Next: Spin up a GPU'), and a 'FREE HOURS LEFT' section that says 'Resets September 30'. The main area is titled 'JUPYTER SERVER fnembhard / dask-cluster'. It shows a 'Jupyter Server' card with a 'stopped' status, a 'Start' button (which is highlighted with a red box), and a 'Metrics' section. The 'Metrics' section includes: Auto Shutoff: 1 hour, Spot Instance: No, SSH URL: (not enabled) (?), App URL: (not enabled), and Docker enabled: No. Below this is a 'Jupyter Lab' button. To the right is a 'Dask Clust...' card with a 'stopped' status, a 'Start' button (also highlighted with a red box), an 'Edit' button, and a 'Delete' button. The 'Metrics' section for the Dask cluster includes: Number of Workers (n\_workers): 3, Number of Worker Processes (nprocs): 1, Number of Worker Threads (nthreads): 2, Scheduler Size: Large - 2 cores - 16 GB RAM, and Spot Instance: No. Below this is a 'Dask Dashboard' button. At the bottom right is a blue circular icon with a white speech bubble.

# Start the cluster

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

The screenshot shows the Saturn Cloud web interface. On the left, a sidebar menu includes options like 'USER fnembhard', 'Create an Organization', 'Resources', 'Secrets', 'Git Repositories', 'Images', 'Enterprise', and 'Get Started Next: Spin up a GPU'. A 'FREE HOURS LEFT' banner at the bottom indicates Resets September 30 and an 'UPGRADE TO PRO' button. The main content area displays a project named 'fnembhard / dask-cluster' with the identifier 'fd8cc8be79064bfc997f04210a154f85'. The 'Overview' tab is selected, showing a 'Jupyter Server' status as 'stopped' with a 'Start' button. Below it, resource details are listed: 'Large - 2 cores - 16 GB RAM - 10Gi Disk'. Metrics include Auto Shutoff: 1 hour, Spot Instance: No, SSH URL: (not enabled), App URL: (not enabled), and Docker enabled: No. A 'Jupyter Lab' button is present. To the right, a 'Dask Cluster' section shows a status bar with 'pending', 'Start', 'Stop', and 'Delete' buttons, and a 'Restart' link. It also lists 'Workers : Large - 2 cores - 16 GB RAM'. Below this, three horizontal progress bars show the status of the 'KUBE CLUSTER', 'SCHEDULER', and 'WORKERS' processes across six stages: Schedule, Provision Hardware, Pull Image, Set Up Environment, Execute Start Script, and Ready. The SCHEDULER bar shows 'Scheduled' and 'Provisioned Hardware' completed, with 'Pulling Image' in progress (labeled '1'). The KUBE CLUSTER bar shows all stages completed. The WORKERS bar shows 'Schedule' completed, 'Provisioning Hardware' in progress (labeled '3'), and 'Pull Image' completed.

# 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.

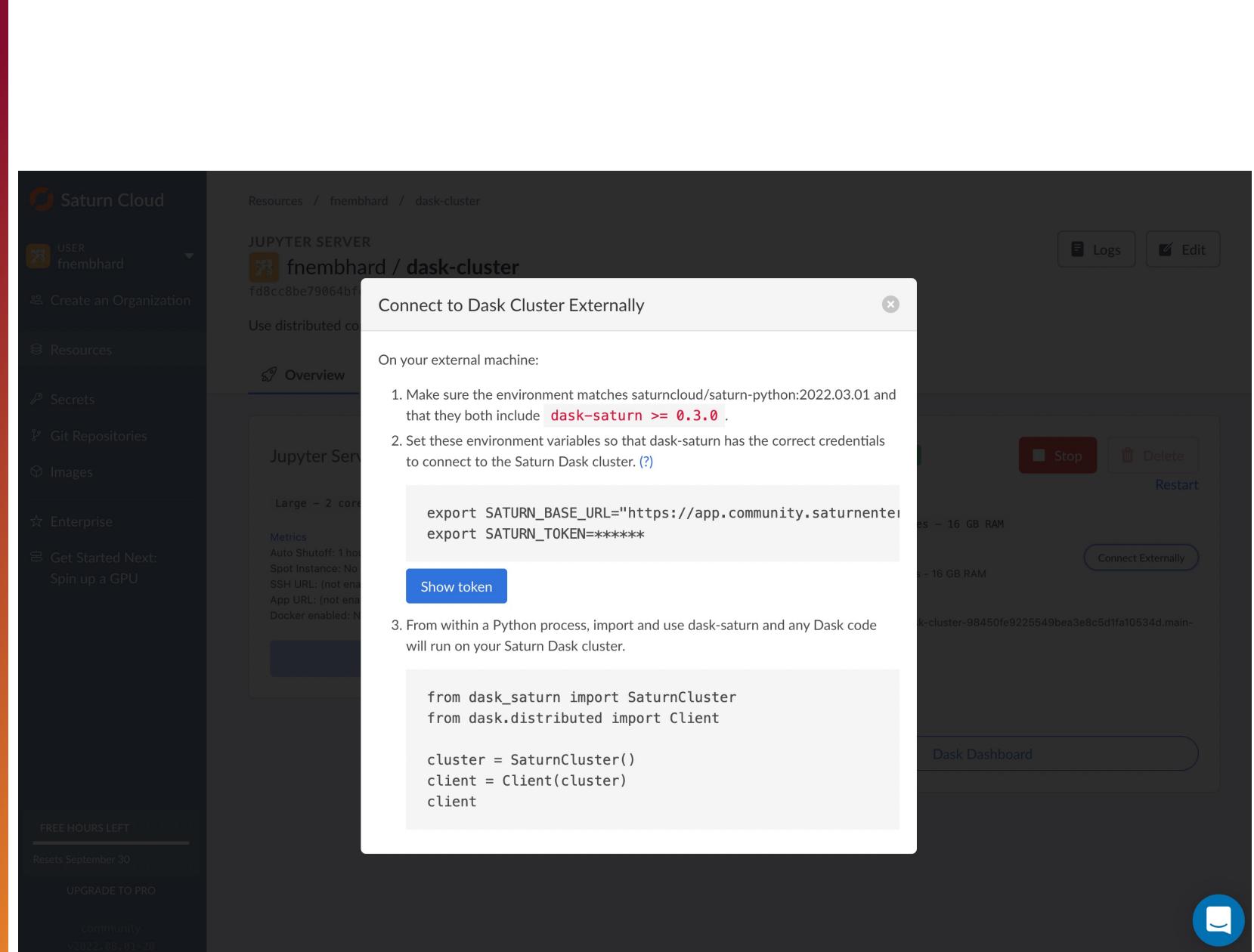
The screenshot shows the Saturn Cloud web interface. On the left is a sidebar with navigation links: USER (fnembhard), Create an Organization, Resources, Secrets, Git Repositories, Images, Enterprise, and Get Started Next: Spin up a GPU. Below the sidebar are banners for 'FREE HOURS LEFT' (Reset September 30) and 'UPGRADE TO PRO' (community v2022.08.01-20). The main content area has a breadcrumb path: Resources / fnembhard / dask-cluster. It displays two sections: 'JUPYTER SERVER' (fnembhard / dask-cluster, fd8cc8be79064bfc997f04210a154f85) and 'Dask Cluster' (running). The Jupyter Server section shows details like 'Large - 2 cores - 16 GB RAM - 10Gi Disk' and metrics. The Dask Cluster section shows workers in various states: Schedule, Provision Hardware, Pulling Image, Setting Up Environment, Executing Start Script, and Ready. A red box highlights the 'Jupyter Lab' button in the Jupyter Server section, and another red box highlights the 'Connect Externally' button in the Dask Cluster section. A blue speech bubble icon is in the bottom right corner.

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.



The screenshot shows the Saturn Cloud web interface. On the left, there's a sidebar with options like 'USER', 'Create an Organization', 'Resources', 'Secrets', 'Git Repositories', 'Images', 'Enterprise', and 'Get Started Next: Spin up a GPU'. Below this is a 'FREE HOURS LEFT' section with a progress bar and a note that it resets on September 30. At the bottom, there are links for 'UPGRADE TO PRO' and 'community'.

The main area shows a 'JUPYTER SERVER' named 'fnembhard / dask-cluster'. It has a status bar indicating 'fd8cc8be79064bf' and 'Use distributed computation'. Below this is an 'Overview' section with a 'Jupyter Server' card showing 'Large - 2 cores' and 'Metrics' (Auto Shutoff: 1 hour, Spot Instance: No, SSH URL: (not enabled), App URL: (not enabled), Docker enabled: No). There are 'Stop', 'Delete', and 'Restart' buttons for the cluster.

A central modal window titled 'Connect to Dask Cluster Externally' contains instructions and code snippets:

- On your external machine:

  1. Make sure the environment matches saturncloud/saturn-python:2022.03.01 and that they both include `dask-saturn >= 0.3.0`.
  2. Set these environment variables so that dask-saturn has the correct credentials to connect to the Saturn Dask cluster. (?)

```
export SATURN_BASE_URL="https://app.community.saturnenterprise.io"
export SATURN_TOKEN=*****
```

[Show token](#)

- 3. From within a Python process, import and use dask-saturn and any Dask code will run on your Saturn Dask cluster.

```
from dask_saturn import SaturnCluster
from dask.distributed import Client

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

At the bottom right of the modal is a 'Dask Dashboard' button. A blue speech bubble icon is located at the bottom right corner of the slide.

# 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:

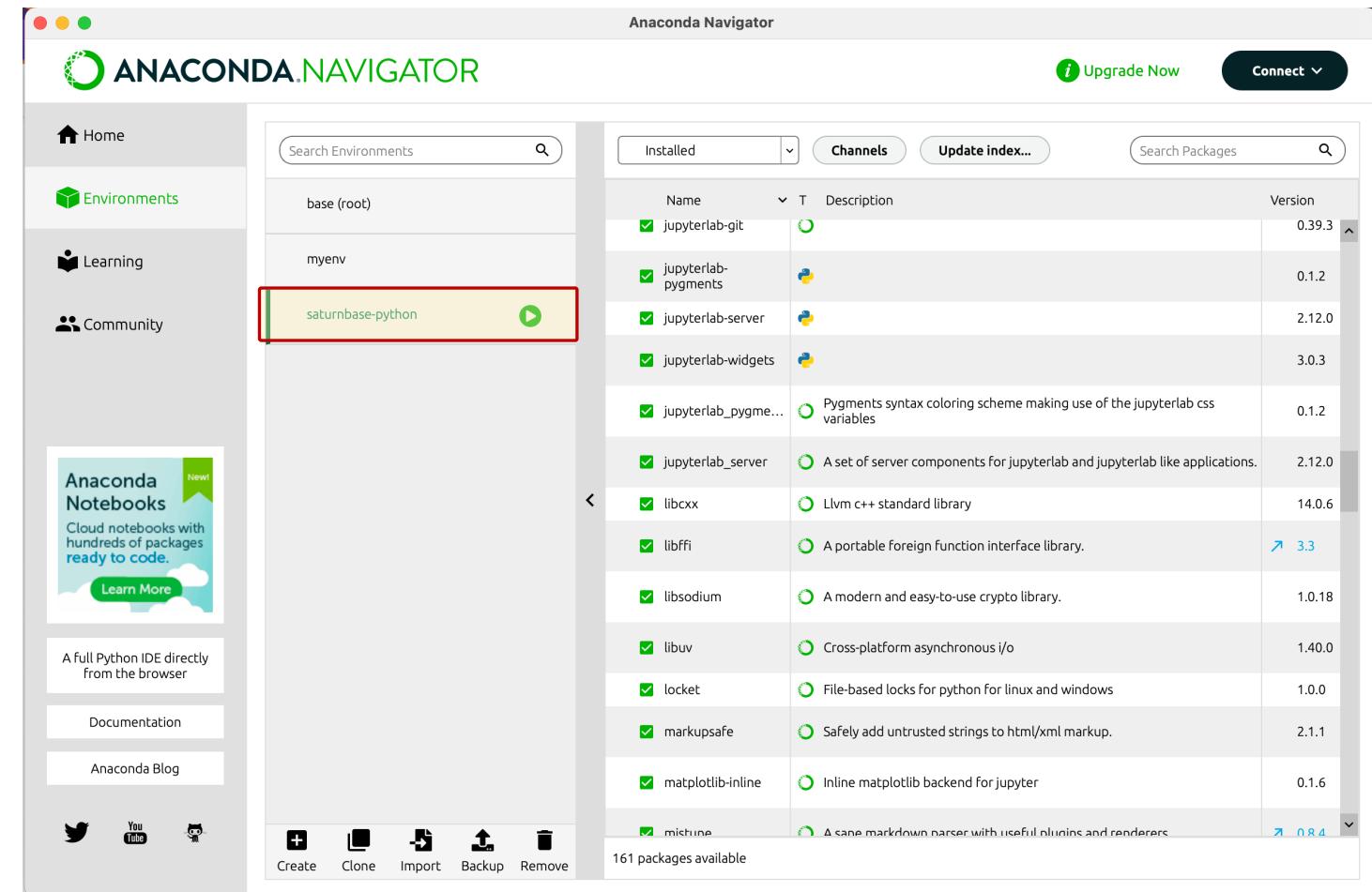
conda env create --name saturnbase-python --file=environment.yml

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

Read more here:

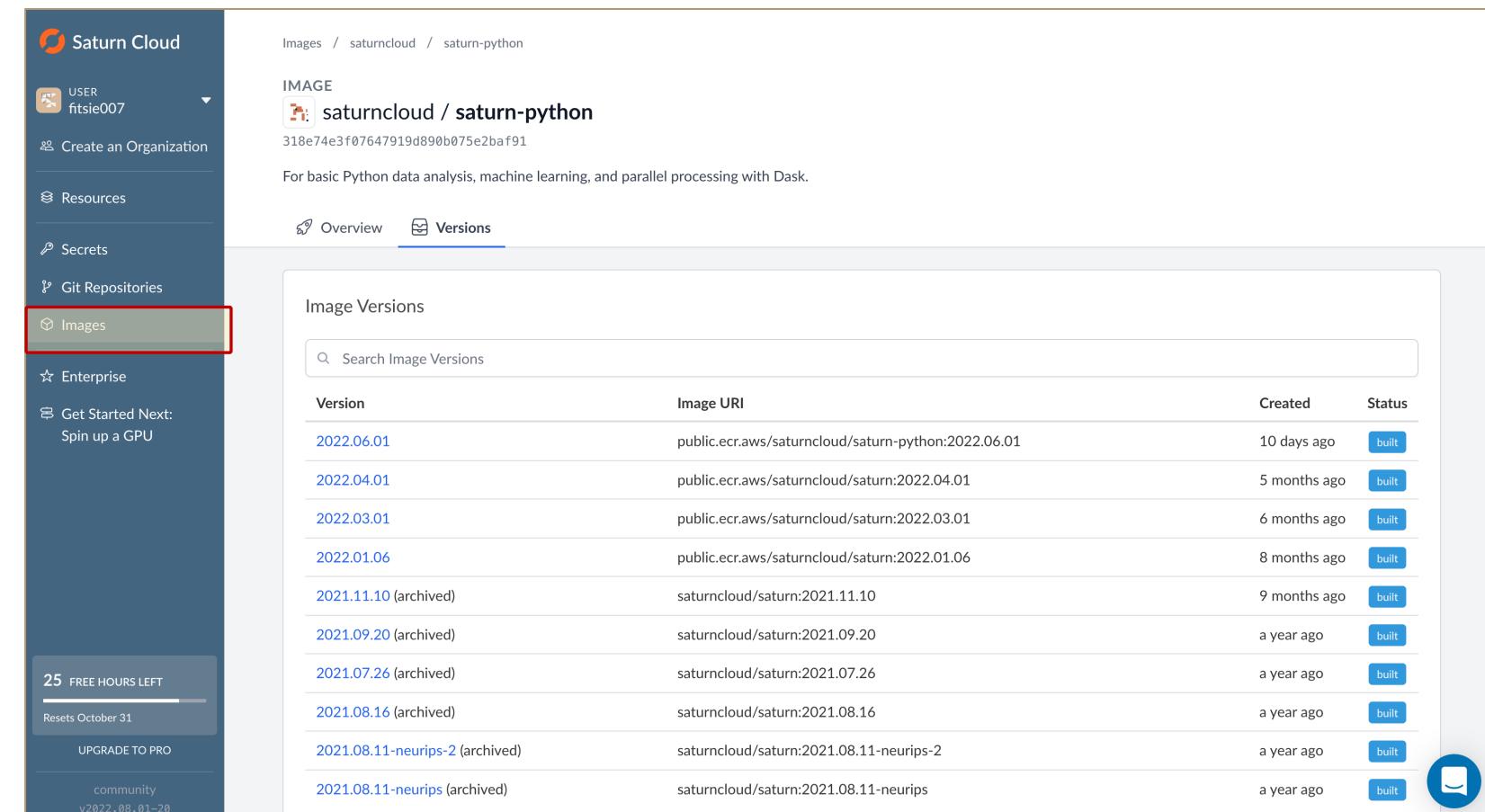
<https://saturncloud.io/docs/using-saturn-cloud/external-connect/>

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.



The screenshot shows the Saturn Cloud web interface. On the left is a sidebar with a user profile for 'fitsie007', options to 'Create an Organization', 'Resources', 'Secrets', 'Git Repositories', and 'Images'. The 'Images' option is highlighted with a red box. Below the sidebar, there's a message about free hours and an upgrade to Pro. The main content area shows the 'saturncloud / saturn-python' image details, including its description: 'For basic Python data analysis, machine learning, and parallel processing with Dask.' Below this are tabs for 'Overview' and 'Versions'. The 'Versions' tab is selected, showing a table of image versions:

Version	Image URI	Created	Status
2022.06.01	public.ecr.aws/saturncloud/saturn-python:2022.06.01	10 days ago	built
2022.04.01	public.ecr.aws/saturncloud/saturn:2022.04.01	5 months ago	built
2022.03.01	public.ecr.aws/saturncloud/saturn:2022.03.01	6 months ago	built
2022.01.06	public.ecr.aws/saturncloud/saturn:2022.01.06	8 months ago	built
2021.11.10 (archived)	saturncloud/saturn:2021.11.10	9 months ago	built
2021.09.20 (archived)	saturncloud/saturn:2021.09.20	a year ago	built
2021.07.26 (archived)	saturncloud/saturn:2021.07.26	a year ago	built
2021.08.16 (archived)	saturncloud/saturn:2021.08.16	a year ago	built
2021.08.11-neurips-2 (archived)	saturncloud/saturn:2021.08.11-neurips-2	a year ago	built
2021.08.11-neurips (archived)	saturncloud/saturn:2021.08.11-neurips	a year ago	built

# 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]: 1 %env SATURN_BASE_URL=https://app.community.saturnenterprise.io
2 %env SATURN_TOKEN=server-IcantLetYouThrowYourselfAwayIcantLetYou
env: SATURN_BASE_URL=https://app.community.saturnenterprise.io
env: SATURN_TOKEN=server-IcantLetYouThrowYourselfAwayIcantLetYou

In [2]: 1 from dask_saturn import SaturnCluster
2 from dask.distributed import Client
3
4 cluster = SaturnCluster()
5 client = Client(cluster)
6 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

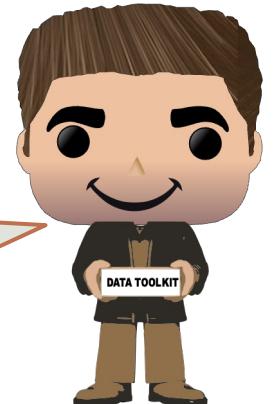
Connection method: Cluster object

Cluster type: SaturnCluster

Dashboard: <https://dc-fnemb-dask-cluster-98450fe9225549bea3e8c5d1fa10534d.community.saturnenterprise.io/>

Cluster Info

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

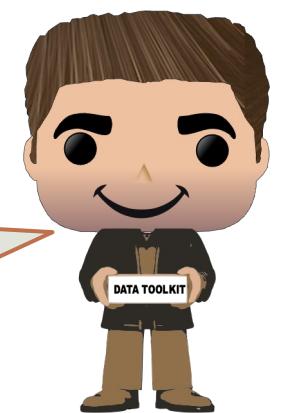


# Connecting to the cluster from Jupyter via Saturn Cloud.

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

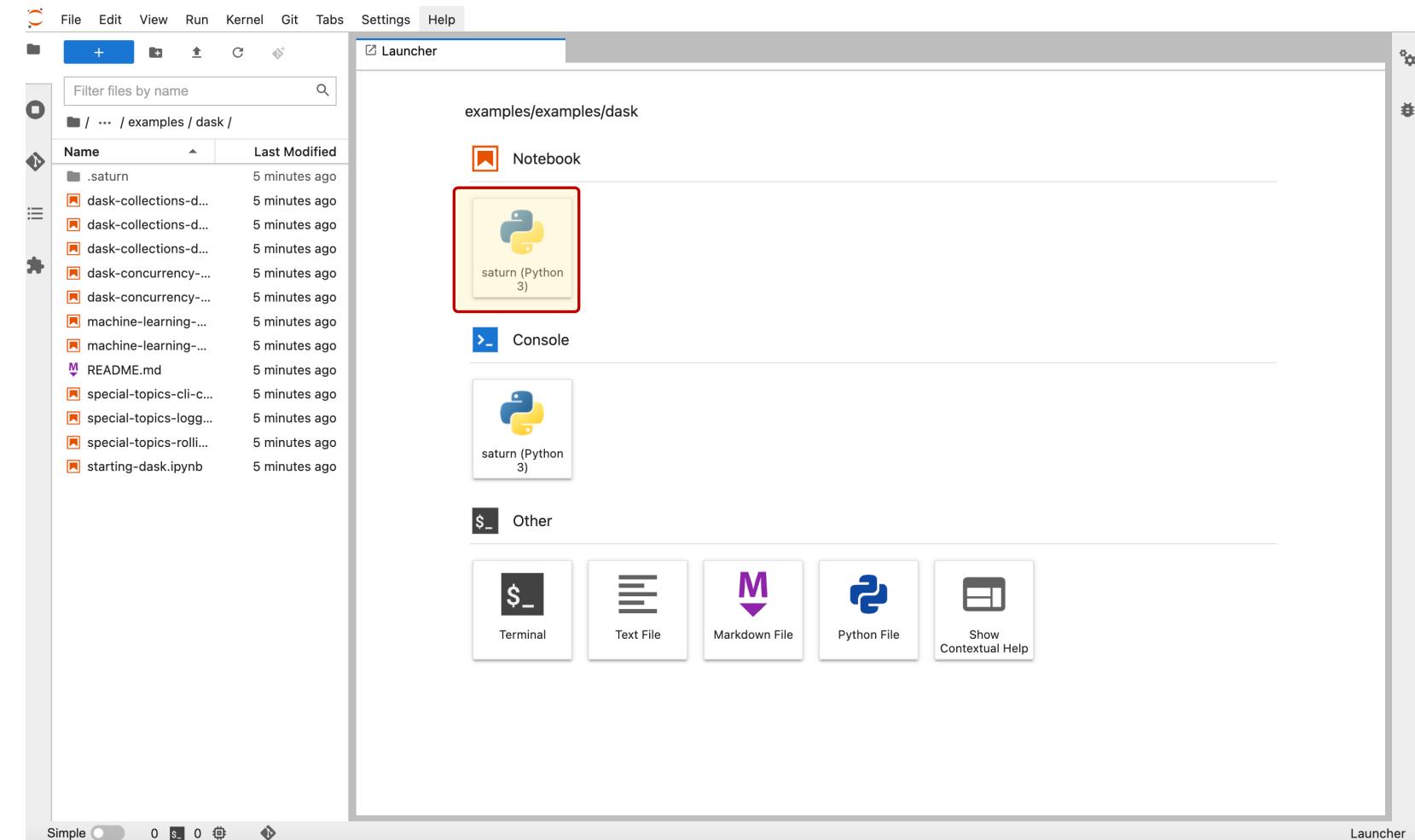
The screenshot shows the Saturn Cloud web interface. On the left is a sidebar with navigation links: USER (fnembhard), Create an Organization, Resources, Secrets, Git Repositories, Images, Enterprise, and Get Started Next: Spin up a GPU. Below this is a section for FREE HOURS LEFT, which resets on September 30. At the bottom are links for UPGRADE TO PRO and community v2022.08.01-20. The main content area has a header for JUPYTER SERVER with the name fnembhard / dask-cluster and a unique identifier fd8cc8be79064bfc997f04210a154f85. It includes a sub-header "Use distributed computing with Dask". Below this are tabs for Overview (which is selected), Environment, Secrets, Git Repos, and Manage. The Overview tab displays two cards: "Jupyter Server" (running) and "Dask Clust...". The Jupyter Server card shows metrics like Large - 2 cores - 16 GB RAM - 10Gi Disk, Auto Shutoff: 1 hour, and SSH URL: (not enabled). The Dask Cluster card shows Workers: Large - 2 cores - 16 GB RAM, Metrics, and a "Connect Externally" button. A blue button labeled "Jupyter Lab" is highlighted with a red border. A blue speech bubble icon is located in the bottom right corner of the main content area.

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



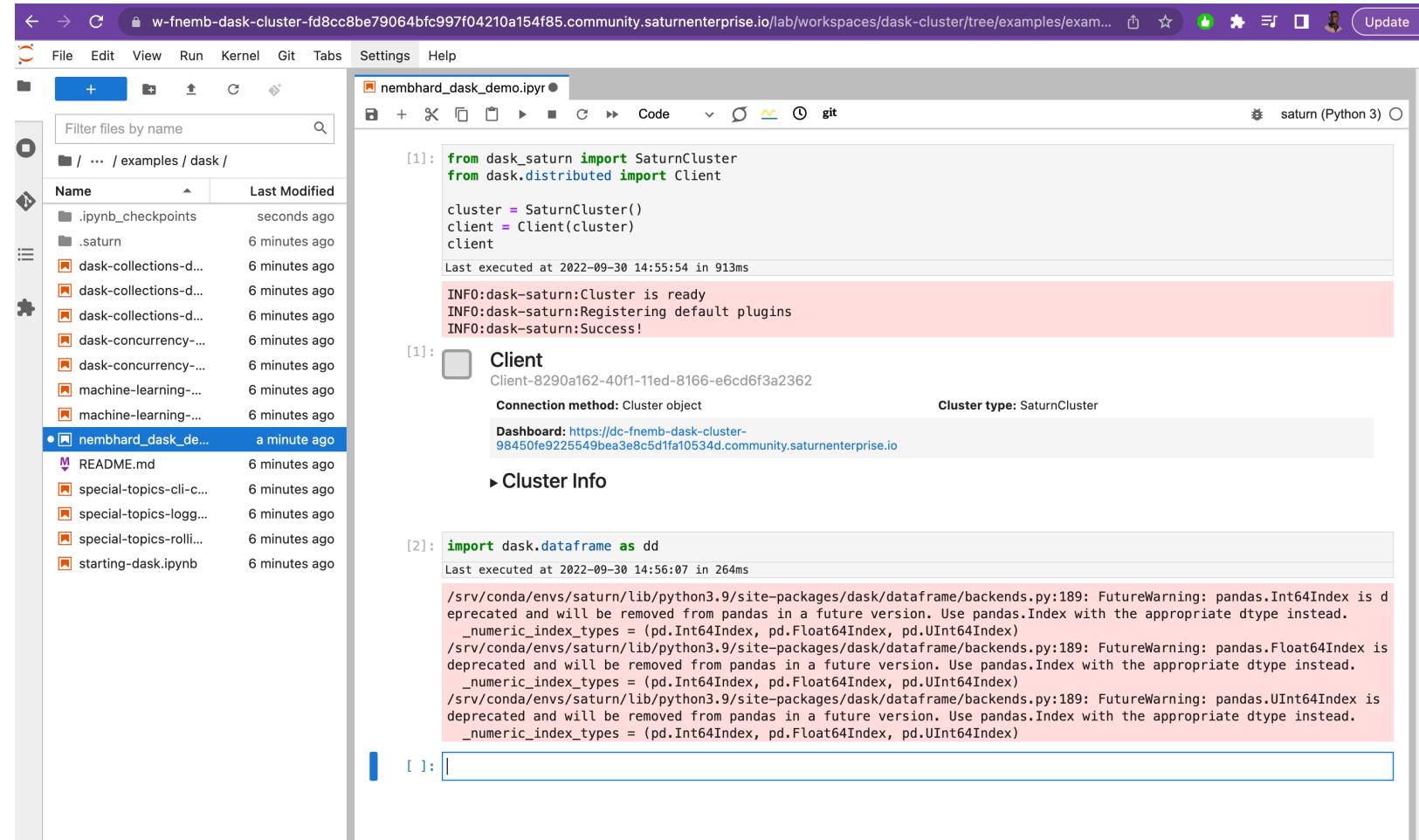
# Connect to and Use the Cluster.

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.



The screenshot shows a Jupyter Lab interface running in a browser window titled "nembhard\_dask\_demo.ipynb". The left sidebar displays a file tree with several Dask-related files and a ".saturn" folder. The main area contains two code cells. Cell [1] shows the initial setup of a Dask cluster using the SaturnCluster and Client modules. It includes log output indicating the cluster is ready, default plugins are registered, and success. Cell [2] imports the dask.dataframe module. Both cells include extensive FutureWarning messages about deprecated pandas index types being removed in future versions. A "Cluster Info" section on the right provides details about the current cluster connection and dashboard URL.

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

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

Last executed at 2022-09-30 14:55:54 in 913ms
INFO:dask-saturn:Cluster is ready
INFO:dask-saturn:Registering default plugins
INFO:dask-saturn:Success!
```

```
[1]: Client
Client-8290a162-40f1-11ed-8166-e6cd6f3a2362
Connection method: Cluster object
Cluster type: SaturnCluster
Dashboard: https://dc-fnemb-dask-cluster-98450fe9225549bea3e8c5d1fa10534d.community.saturnenterprise.io
```

```
[2]: import dask.dataframe as dd
Last executed at 2022-09-30 14:56:07 in 264ms

/srv/conda/envs/saturn/lib/python3.9/site-packages/dask/dataframe/backends.py:189: FutureWarning: pandas.Int64Index is deprecated and will be removed from pandas in a future version. Use pandas.Index with the appropriate dtype instead.
    _numeric_index_types = (pd.Int64Index, pd.Float64Index, pd.UInt64Index)
/srv/conda/envs/saturn/lib/python3.9/site-packages/dask/dataframe/backends.py:189: FutureWarning: pandas.Float64Index is deprecated and will be removed from pandas in a future version. Use pandas.Index with the appropriate dtype instead.
    _numeric_index_types = (pd.Int64Index, pd.Float64Index, pd.UInt64Index)
/srv/conda/envs/saturn/lib/python3.9/site-packages/dask/dataframe/backends.py:189: FutureWarning: pandas.UInt64Index is deprecated and will be removed from pandas in a future version. Use pandas.Index with the appropriate dtype instead.
    _numeric_index_types = (pd.Int64Index, pd.Float64Index, pd.UInt64Index)
```

## **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>

The screenshot shows the 'Global Summary of the Month (GSOM), Version 1' search interface. The URL in the browser is `ncei.noaa.gov/access/search/data-search/global-summary-of-the-month?bbox=40.965,-74.257,40.465,-73.757&startDate=2014-01-01T00:00:00&endDate=2021-12-31T23:59:59`. The page features a header with the NOAA/NCEI logo and navigation links for Home, Access, Search, Dataset Search, Data Search, Order Status, Help, Guide, and Cart (0). A search bar is also present. The main content area is titled 'Global Summary of the Month (GSOM), Version 1' with a 'Clear Search' button. It includes three main search sections: 'What' (Data Types, Show List button), 'Where' (Location input, Find Location Using Map button, coordinates input), and 'When' (Date range input, Select Date Range checked, date pickers for year, month, and day). Below these is a 'Station Search' section with an input field for 'Ex: Airport'.

# 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>

The screenshot shows the National Weather Service Climate page for New York, NY. At the top, there's a banner about Hurricane Ian. Below it, the 'Climate' section is selected in the navigation bar. The main content area is titled 'NOWData - NOAA Online Weather Data' and includes four tabs: NOWData (selected), Observed Weather, Climate Prediction and Variability, Local Data/Records, and Climate Resources. The NOWData tab shows a form for selecting location (New York, NY is selected), product (Monthly summarized data is selected), and variable (Avg temp). It also includes a year range (2000 - 2022) and a summary dropdown set to 'Mean'. A 'Product Description' box explains monthly summarized data calculations. At the bottom, it mentions the Applied Climate Information System (ACIS) and its partners.

# 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  
   
[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																																																																																																																																																																																																																																																																																																																																																															
<b>NOWData - NOAA Online Weather Data</b> <div style="border: 1px solid black; padding: 5px;"> <b>Monthly Mean Avg Temperature for NY CITY CENTRAL PARK, NY</b>  <small>Click column heading to sort ascending, click again to sort descending.</small> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th>Year</th> <th>Jan</th> <th>Feb</th> <th>Mar</th> <th>Apr</th> <th>May</th> <th>Jun</th> <th>Jul</th> <th>Aug</th> <th>Sep</th> <th>Oct</th> <th>Nov</th> <th>Dec</th> <th>Annual</th> </tr> </thead> <tbody> <tr><td>2000</td><td>31.3</td><td>37.3</td><td>47.2</td><td>51.0</td><td>63.5</td><td>71.3</td><td>72.3</td><td>72.4</td><td>66.0</td><td>57.0</td><td>45.3</td><td>31.1</td><td>53.8</td></tr> <tr><td>2001</td><td>33.6</td><td>35.9</td><td>39.6</td><td>53.9</td><td>63.6</td><td>72.9</td><td>73.1</td><td>78.7</td><td>67.7</td><td>58.5</td><td>52.7</td><td>44.1</td><td>56.2</td></tr> <tr><td>2002</td><td>39.9</td><td>40.6</td><td>44.1</td><td>56.1</td><td>60.7</td><td>71.5</td><td>78.8</td><td>77.7</td><td>70.2</td><td>55.2</td><td>46.0</td><td>36.0</td><td>56.4</td></tr> <tr><td>2003</td><td>27.5</td><td>30.1</td><td>43.1</td><td>49.8</td><td>58.7</td><td>68.4</td><td>75.8</td><td>76.7</td><td>67.9</td><td>55.1</td><td>50.0</td><td>37.6</td><td>53.4</td></tr> <tr><td>2004</td><td>24.7</td><td>35.0</td><td>43.5</td><td>53.6</td><td>65.2</td><td>71.2</td><td>74.5</td><td>74.2</td><td>69.3</td><td>56.0</td><td>48.2</td><td>38.4</td><td>54.5</td></tr> <tr><td>2005</td><td>31.3</td><td>36.5</td><td>39.5</td><td>55.1</td><td>58.9</td><td>74.0</td><td>77.5</td><td>79.7</td><td>73.3</td><td>57.9</td><td>49.6</td><td>35.3</td><td>55.7</td></tr> <tr><td>2006</td><td>40.9</td><td>35.7</td><td>43.1</td><td>55.7</td><td>63.1</td><td>71.0</td><td>77.9</td><td>75.8</td><td>66.6</td><td>56.2</td><td>51.9</td><td>43.6</td><td>56.8</td></tr> <tr><td>2007</td><td>37.5</td><td>28.3</td><td>42.2</td><td>50.3</td><td>65.2</td><td>71.4</td><td>75.0</td><td>74.0</td><td>70.3</td><td>63.6</td><td>45.4</td><td>37.0</td><td>55.0</td></tr> <tr><td>2008</td><td>36.5</td><td>35.8</td><td>42.6</td><td>55.0</td><td>60.1</td><td>74.0</td><td>78.4</td><td>73.8</td><td>68.8</td><td>55.1</td><td>45.9</td><td>38.1</td><td>55.3</td></tr> <tr><td>2009</td><td>27.9</td><td>36.7</td><td>42.4</td><td>54.5</td><td>62.5</td><td>67.5</td><td>72.7</td><td>75.7</td><td>66.3</td><td>55.0</td><td>51.1</td><td>35.9</td><td>54.0</td></tr> <tr><td>2010</td><td>32.5</td><td>33.1</td><td>48.2</td><td>57.9</td><td>65.3</td><td>74.7</td><td>81.3</td><td>77.4</td><td>71.1</td><td>58.1</td><td>47.9</td><td>32.8</td><td>56.7</td></tr> <tr><td>2011</td><td>29.7</td><td>36.0</td><td>42.3</td><td>54.3</td><td>64.5</td><td>72.3</td><td>80.2</td><td>75.3</td><td>70.0</td><td>57.1</td><td>51.9</td><td>43.3</td><td>56.4</td></tr> <tr><td>2012</td><td>37.3</td><td>40.9</td><td>50.9</td><td>54.8</td><td>65.1</td><td>71.0</td><td>78.8</td><td>76.7</td><td>68.8</td><td>58.0</td><td>43.9</td><td>41.5</td><td>57.3</td></tr> <tr><td>2013</td><td>35.1</td><td>33.9</td><td>40.1</td><td>53.0</td><td>62.8</td><td>72.7</td><td>79.8</td><td>74.6</td><td>67.9</td><td>60.2</td><td>45.3</td><td>38.5</td><td>55.3</td></tr> <tr><td>2014</td><td>28.6</td><td>31.6</td><td>37.7</td><td>52.3</td><td>64.0</td><td>72.5</td><td>76.1</td><td>74.5</td><td>69.7</td><td>59.6</td><td>45.3</td><td>40.5</td><td>54.4</td></tr> <tr><td>2015</td><td>29.9</td><td>23.9</td><td>38.1</td><td>54.3</td><td>68.5</td><td>71.2</td><td>78.8</td><td>79.0</td><td>74.5</td><td>58.0</td><td>52.8</td><td>50.8</td><td>56.7</td></tr> <tr><td>2016</td><td>34.5</td><td>37.7</td><td>48.9</td><td>53.3</td><td>62.8</td><td>72.3</td><td>78.7</td><td>79.2</td><td>71.8</td><td>58.8</td><td>49.8</td><td>38.3</td><td>57.2</td></tr> <tr><td>2017</td><td>38.0</td><td>41.6</td><td>39.2</td><td>57.2</td><td>61.1</td><td>72.0</td><td>76.8</td><td>74.0</td><td>70.5</td><td>64.1</td><td>46.6</td><td>35.0</td><td>56.3</td></tr> <tr><td>2018</td><td>31.7</td><td>42.0</td><td>40.1</td><td>49.5</td><td>66.9</td><td>71.7</td><td>77.6</td><td>78.1</td><td>70.7</td><td>57.7</td><td>44.4</td><td>40.1</td><td>55.9</td></tr> <tr><td>2019</td><td>32.5</td><td>36.2</td><td>41.7</td><td>55.5</td><td>62.2</td><td>71.7</td><td>79.6</td><td>75.5</td><td>70.4</td><td>59.9</td><td>43.9</td><td>38.3</td><td>55.6</td></tr> <tr><td>2020</td><td>39.1</td><td>40.1</td><td>48.0</td><td>50.4</td><td>60.3</td><td>73.7</td><td>80.0</td><td>76.9</td><td>68.8</td><td>57.9</td><td>53.0</td><td>39.2</td><td>57.3</td></tr> <tr><td>2021</td><td>34.8</td><td>34.2</td><td>45.8</td><td>54.6</td><td>62.9</td><td>74.3</td><td>76.0</td><td>77.5</td><td>70.3</td><td>62.0</td><td>46.2</td><td>43.8</td><td>56.9</td></tr> <tr><td>2022</td><td>30.3</td><td>37.3</td><td>45.3</td><td>52.8</td><td>64.0</td><td>71.4</td><td>79.5</td><td>79.3</td><td>70.0</td><td>M</td><td>M</td><td>M</td><td>58.9</td></tr> <tr> <td></td><td><b>Mean</b></td><td>33.3</td><td>35.7</td><td>43.2</td><td>53.7</td><td>63.1</td><td>71.9</td><td>77.4</td><td>76.4</td><td>69.6</td><td>58.2</td><td>48.0</td><td>39.1</td><td>55.9</td></tr> </tbody> </table> </div>					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		<b>Mean</b>	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
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																																																																																																																																																																																																																																																																																																																																																						
	<b>Mean</b>	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																																																																																																																																																																																																																																																																																																																																																					

# 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>

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

 NOAA NATIONAL CENTERS FOR ENVIRONMENTAL INFORMATION

Home Climate Information Data Access Contact About Search Datasets Search Tool Mapping Tool Data Tools help

Climate Data Online: Web Services Documentation

Getting Started Datasets Data Categories Data Types Location Categories Locations Stations Data

Data

The data endpoint is used for actually fetching the data.

Endpoints

PATH	DESCRIPTION
<a href="https://www.ncdc.noaa.gov/cdo-web/api/v2/data?datasetid=YOUGDATASETID">https://www.ncdc.noaa.gov/cdo-web/api/v2/data?datasetid=YOUGDATASETID</a>	Requires exactly one dataset id. Used to fetch data.

Additional Parameters

PARAMETER	EXAMPLE	DESCRIPTION
datasetid	GSOM	Required. Accepts a single valid dataset id. Data returned will be from the dataset specified.
datatypeid	ACMR	Optional. Accepts a valid data type id or a chain of data type ids separated by ampersands. Data returned will contain all of the data type(s) specified.
locationid	FIPB:37	Optional. Accepts a valid location id or a chain of location ids separated by ampersands. Data returned will contain data for the location(s) specified.
stationid	GHCND:US1NCBC0005	Optional. Accepts a valid station id or a chain of station ids separated by ampersands. Data returned will contain data for the station(s) specified.
startdate	1970-10-03	Required. Accepts valid ISO formatted date (YYYY-MM-DD) or date time (YYYY-MM-DDThhmmss). Data returned will be after the specified date. Annual and Monthly data will be limited to a ten year range while all other data will be limited to a one year range.
enddate	2012-09-10	Required. Accepts valid ISO formatted date (YYYY-MM-DD) or date time (YYYY-MM-DDThhmmss). Data returned will be before the specified date. Annual and Monthly data will be limited to a ten year range while all other data will be limited to a one year range.
units	metric	Optional. Accepts the literal strings "standard" or "metric". Data will be scaled and converted to the specified units. If a unit is not provided then no scaling nor conversion will take place.
sortfield	name	Optional. The field to sort results by. Supports id, name, mindata, maxdate, and datacoverage fields.
sortorder	desc	Optional. Which order to sort by, asc or desc. Defaults to asc.
limit	42	Optional. Defaults to 25, limits the number of results in the response. Maximum is 1000.
offset	24	Optional. Defaults to 0, used to offset the resultlist. The example would begin with record 24.
includemetadata	false	Optional. Defaults to true, used to improve response time by preventing the calculation of result metadata.

Examples

```
will display example results from
low to use the following header
a thin queries and must be in the header
res) for zip code 2801, May 1st of 2010
2/data?datasetid=GHCND:locationid=ZIP:2801&startdate=2010-05-01&enddate=2010-05-01
on 15 Minuto for COOP station 010008, for May of 2010 with metric units
2/data?datasetid=PRCP_15min&stationid=COOP:010008&units=metric&startdate=2010-05-01&enddate=2010-05-31
try of the Month) for GHCN station USC00010008, for May of 2010 with standard units
2/data?datasetid=GSOM&stationid=GHCND:USC00010008&units=standard&startdate=2010-05-01&enddate=2010-05-31
```

# 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/external-connect/external_connect/)
- <https://saturncloud.io/docs/using-saturn-cloud/secrets/>