

---

## Guide: PySpark on the UCloud HPC Cluster

---

Thorvald Sørensen, thso@itu.dk

February 9, 2024

To handle big datasets, you can use a Spark Cluster.

UCloud/DeiC provides you with a platform where you can store files and run apps. You need to setup your own very own Spark Cluster (what a luxury! Normally you have to share) and JupyterLab “app”. To use the app, you need to specify which computing resources it should use and create a job. To pay for the resources you can use your 1000 DKK free credit and select a fitting amount of hours your job should run.

### NOTE ON RE-STARTING EVERYTHING

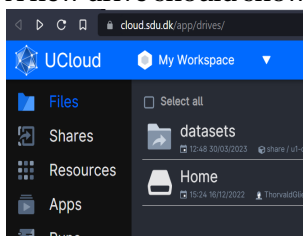
You always need to start the Spark Cluster job before you start the JupyterLab job. When starting the Jupyter Lab, you always need to connect the job to the cluster job, by specifying the most recent job id of the spark cluster, by selecting current running job. Verify that the current running spark cluster job id is the same you see before submitting the spark cluster job.

## 1 CREATE A SPARK CLUSTER JOB ON DEIC INTERACTIVE HPC

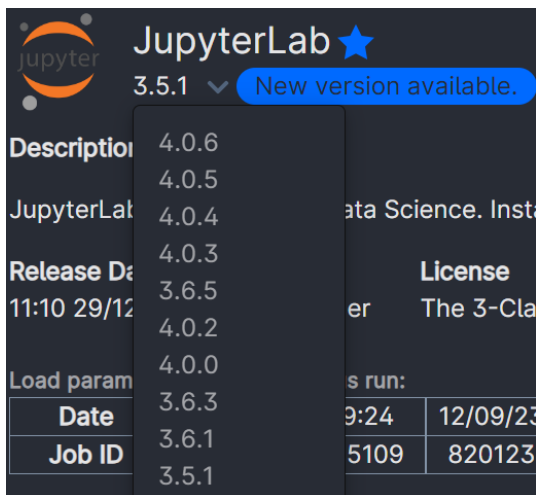
First we need to request computing resources from the cluster and get the data available on the cluster.

- 1) Log in to [DeiC interactive HPC](#) with ITU account (Wayf)
- 2) Add the “datasets” to your [drive](#) (containing the yelp dataset) using an invite link from a TA. A current link is on the Learnit page for this assignment.

A new drive should show up:



- 3) Find [Spark Cluster](#) under the “apps” tab
- 4) Click the version number and a drop-down menu appears, select version [3.3.1](#) and click the star icon to easily access the right version under apps next time you work on the assignment:



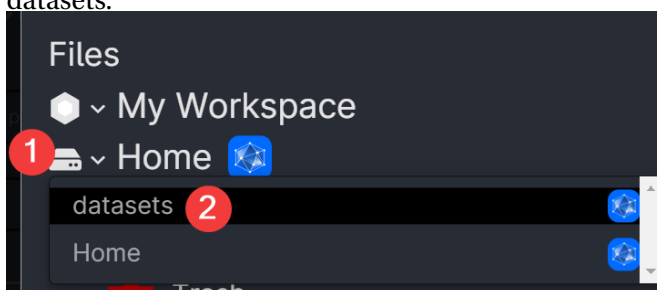
### IMPORTANT

Use Spark Cluster version 3.3.1 is tested by TAs, we had problems with newer versions (not all combinations of Spark Cluster and JupyterLab are compatible).

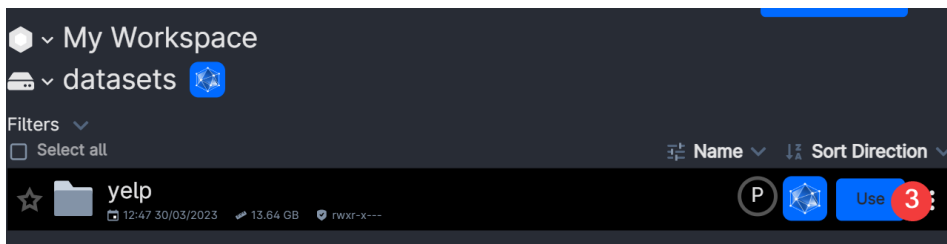
- 5) Create a new job with below configuration:

The screenshot shows the Spark Cluster configuration form. The 'Job name' field has a placeholder 'Example: Run with parameters XYZ'. The 'Hours' field is set to '2'. The 'Number of nodes' field is set to '3'. The 'Machine type' is 'u1-standard-4'. The 'vCPU' is '4 (Intel Xeon Gold 6130)', 'Memory (GB)' is '24', 'GPU' is 'None', and 'Price' is '0,33 DKK/hour'. The 'Mandatory Parameters' section has an 'Input folder' field set to '/datasets/yelp'. A note below states 'folder available on all the cluster nodes'.

- 6) To add the yelp folder as input folder, click the field, then click the drive icon, and select datasets.



- 7) Then click "use" on the yelp folder.



#### Note

You only have 1000 DKK free credit. You pay per hour and for hardware resources. Above configuration can be changed to use fewer resources (less hours CPUs, memory, nodes..) to pay less. We recommend watching your resource usage closely, so you do not risk running out of credit before you have made the assignment.

- 8) Submit the job (may take a few minutes depending on how many people are using the cluster).

#### Note

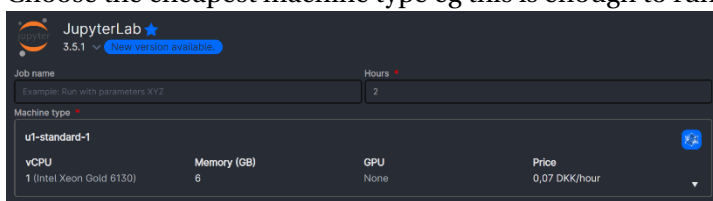
You should wait for the cluster to have started before starting the JupyterLab, since you cannot connect the two jobs otherwise.

## 2 CREATE A JUPYTERLAB JOB ON DEIC INTERACTIVE HPC

Now that you have a Spark Cluster, you just need a place to write code that can be executed on the remote cluster. JupyterLab is an extension to Jupyter notebooks, where you can have multiple “tabs” and “windows” with different files at the same time. All the code you will write will be run “remotely” on the Cluster and not locally on the notebook (as you are used to from your laptop)

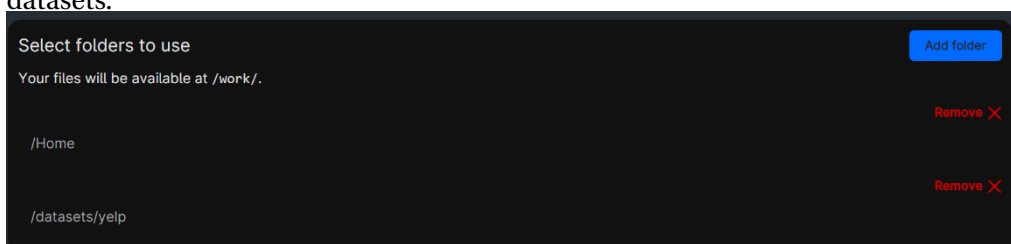
- 1) Wait until your Spark Cluster job is running (no longer in queue)
- 2) Find (and click star) the [JupyterLab app 3.5.1](#)

Choose the cheapest machine type eg this is enough to run JupyterLab:



(No heavy work is being done on this machine. The 3 spark cluster machines will do the heavy calculations)

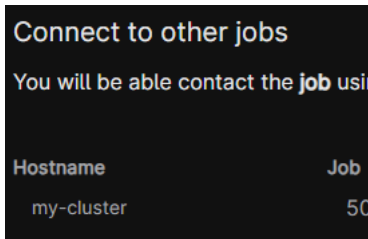
- 3) Select folders that will be available in your Jupyter Lab app. You can simply specify the root of your folders (/Home). Additionally, you need to specify the yelp folder with the datasets.



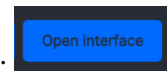
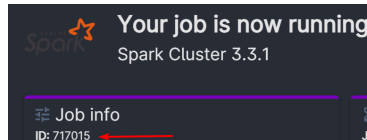
- 4) Connect the JupyterLab to the Spark Cluster job. Set hostname to **my-cluster**.

#### Note

Note that the Job ID will change every time you start the Spark Cluster job, thus you need to click the job id to re-connect to the right id every time you start the job.



- 5) Submit the job. Wait some minutes.



- 6) When the job is running: Click:

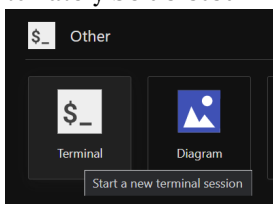
### 3 GETTING THE TEMPLATE NOTEBOOK ON UCloud

For this assignment you need to submit your code with the answer to queries on ITU's GitHub. Upload the Jupyter notebook template to get started. By doing so, you will learn about git, a useful industry skill! Firstly you need to create a repository, a place to store your code for the assignment.

#### Note

If you are new to Git, please take a look at the [missing semester course](#) and do exercise 1-7. A side learning outcome of these assignments is to learn Git, a useful industry skill.

1. Use the repository template to get started and create your own repository.
2. In jupyter lab, you can clone your repository using a terminal. If you use the git GUI menu, make sure to move the repository folder to your UCloud home drive. It will unfortunately be **deleted** when the jupyter lab app shuts down, unless you move it.



- 3.
4. Clone using your repository URL under the folder you selected when you started the JupyterLab job, could be to avoid losing your work.

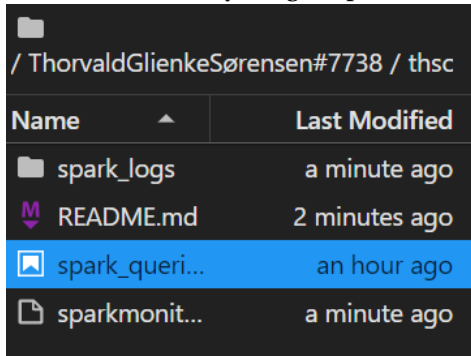
```
1 git clone <insert_your_repository_url_here> Home/<some_folder_you_decide>
```

#### Note

Do not clone this folder to your local computers disk. Your local computer does not have access to the spark cluster. All the code should be stored and run from the UCloud JupyterLab session. Also note to write "Home" and not "home". That folder contains all *your* files on the UCloud drive.

5. Optional: Use SSH for authentication instead of passwords. Requires setting up a new ssh-key in your *persistent* /home on Ucloud and adding the public key to ITU's github.

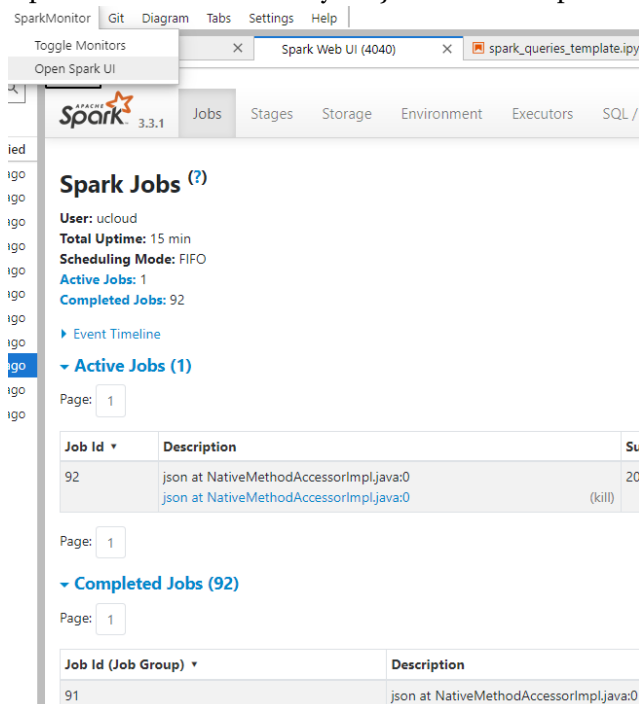
6. A new folder with your git repo has been created.



Name	Last Modified
spark_logs	a minute ago
README.md	2 minutes ago
spark_queri...	an hour ago
sparkmonit...	a minute ago

7. Open the notebook template to get started with the assignment!

Optional: Check status of your jobs with the Spark Web UI:



SparkMonitor Git Diagram Tabs Settings Help

Toggle Monitors X Spark Web UI (4040) X spark\_queries\_template.ipyn

Open Spark UI

Spark 3.3.1 Jobs Stages Storage Environment Executors SQL /

### Spark Jobs (?)

User: ucloud  
Total Uptime: 15 min  
Scheduling Mode: FIFO  
Active Jobs: 1  
Completed Jobs: 92

Event Timeline

Active Jobs (1)

Page: 1

Job Id	Description	Status
92	json at NativeMethodAccessorImpl.java:0 json at NativeMethodAccessorImpl.java:0	20: (kill)

Page: 1

Completed Jobs (92)

Page: 1

Job Id (Job Group)	Description
91	json at NativeMethodAccessorImpl.java:0

### Why did my job shutdown?

After X hours, your job will shutdown and then you need to start it again, unless you extend it.

### Do I need to fill out above configuration parameters every time?

You can load parameters from a previous run by clicking on the run ID table, to easily re-use a configuration e.g. folders to use.

### Why are my queries taking so long to execute?

Perhaps you need more cores. Try to shutdown your jupyter lab and spark cluster, then start the spark cluster with more resources.

### What is UCloud?

A cloud provider company.

### What is DeiC?

The Danish e-infrastructure Cooperation (DeiC) coordinates Danish digital infrastructure as an umbrella for the eight Danish universities to ensure delivery of computing, storage and network infrastructure to Danish research, teaching and innovation.

### Additional resources:

<https://docs.cloud.sdu.dk/Apps/jupyter-lab.html#cluster-deployment>

<https://docs.cloud.sdu.dk/Apps/spark-cluster.html>