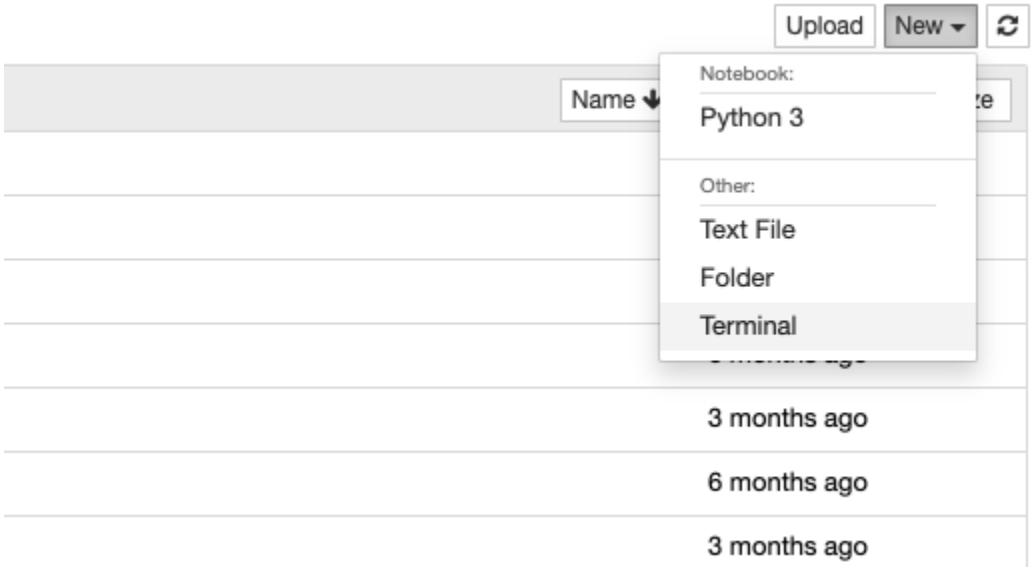
Starting a spark cluster

1. Login to datahub.ucsd.edu
2. Select the “DSC 291” environment and click the “Launch Environment” button

3. Once your notebook server is running, select “New > Terminal”



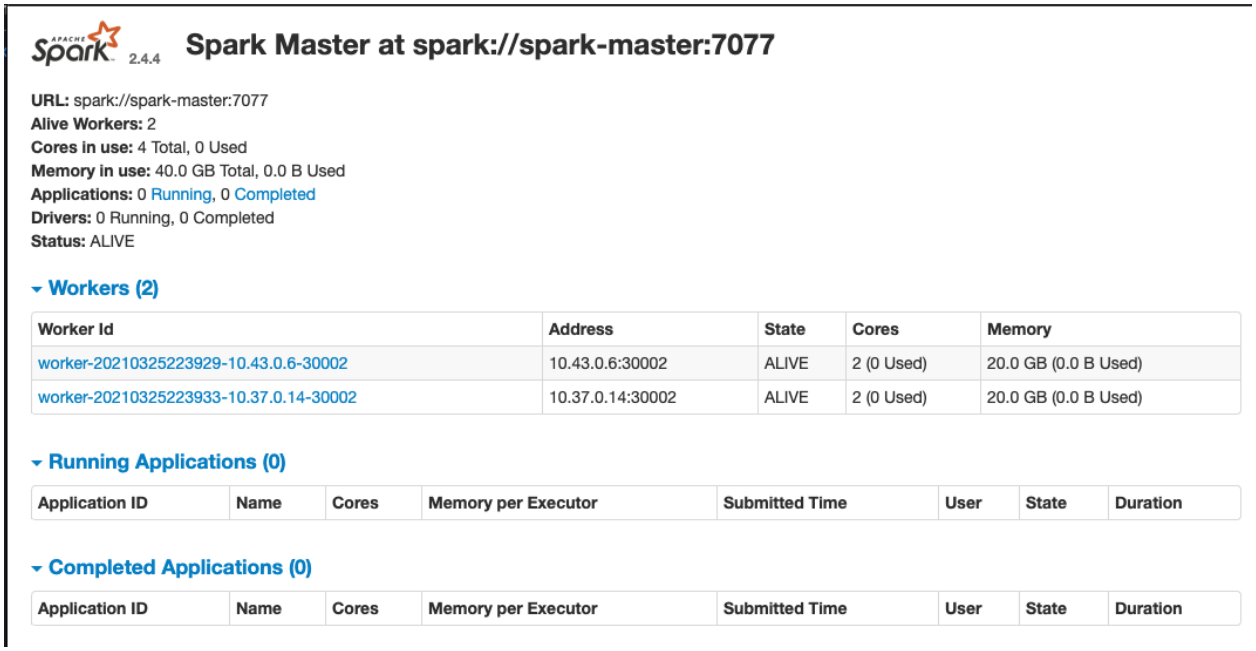
4. Switch to the terminal tab in your web browser
5. Once at the terminal, run the command “**ssh <username>@dsmlp-login**”. Replace <username> with your username.
6. Download the scripts for your course by running the command “**git clone** https://github.com/ucsd-ets/dsc291-spark-cluster.git”
7. You should now see a directory called dsc291-spark-container within your terminal (run the command “**ls**” to list the contents of the directory).
8. Change into the directory by running the command “**cd dsc291-spark-container**”
9. To start your spark cluster, run the command “**./cluster-manager.sh create**”. You should see output on your screen like the example below:

```
-bash-4.2$ ./cluster-manager.sh create
NAME                            READY   STATUS              RESTARTS   AGE
dsmlp-jupyter-wuykimpang        1/1     Running             0          9m34s
spark-master-766b4d8588-j8bt5   0/1     ContainerCreating   0          2s
spark-worker-65588c4fcb-gr2t5   0/1     ContainerCreating   0          2s
NAME                            READY   STATUS    RESTARTS   AGE
dsmlp-jupyter-wuykimpang        1/1     Running   0          9m39s
spark-master-766b4d8588-j8bt5   1/1     Running   0          7s
spark-worker-65588c4fcb-ghbrq   1/1     Running   0          4s
spark-worker-65588c4fcb-gr2t5   1/1     Running   0          7s

======================================================================
=> Successfully initiated the Spark cluster
=> Next create a SSH tunnel from your personal computer using the following command:
       ssh -N -L 127.0.0.1:8080:127.0.0.1:33783 -L 127.0.0.1:4040:127.0.0.1:36578 wuykimpang@dsmlp-login.ucsd.edu

=> Link to Spark cluster manager UI: http://127.0.0.1:8080
=> Link to Spark job UI: http://127.0.0.1:4040
======================================================================
```

10. Open a new terminal on your local computer (**not in Datahub**) and paste the generated ssh command into it (command right below “Next create a SSH tunnel…”). This will open a tunnel between datahub servers and your local computer. Leave this terminal open to keep the tunnel open. ***Note: make sure you’re connected to UCSD’s VPN***
11. Open a new tab in your browser and navigate to http://127.0.0.1:8080. The Apache Spark dashboard will be there. See example below:



12. You can now close the terminal. ***Note: your cluster will only be active for 3 hours. You’ll have to recreate it starting from step 3 in case it shuts down while you’re working with it.***
13. You can now start accessing the generated spark cluster within your jupyter server on Datahub. Please speak with your Instructor or TA about how to access it.

Common problems

Different output at step 9

If you see output like the following:

```
-bash-4.2$ ./cluster-manager.sh create
Error from server (AlreadyExists): error when creating "STDIN": deployments.extensions "spark-master"
 already exists Error from server (AlreadyExists): error when creating "STDIN": services "spark-maste
r" already exists Error from server (AlreadyExists): error when creating "STDIN": deployments.extensi
ons "spark-worker" already exists Error from server (AlreadyExists): error when creating "STDIN": ser
vices "spark-worker" already exists
-bash-4.2$
```

That means that your spark cluster is already running and you may start using it. You can also recreate the cluster by running the command "**./cluster-manager.sh delete**" and then the command "**./cluster-manager.sh create**".