Running Spark + MLlib locally (MacOS):

Installation/setup notes:

1) Install pyspark

```
% pip3 install pyspark
Collecting pyspark
  Downloading pyspark-3.0.2.tar.gz (204.8 MB)
     |████████████████████████████████| 204.8 MB 6.0 MB/s
Collecting py4j==0.10.9
  Downloading py4j-0.10.9-py2.py3-none-any.whl (198 kB)
     |████████████████████████████████| 198 kB 2.2 MB/s
Using legacy 'setup.py install' for pyspark, since package 'wheel' is not installed.
Installing collected packages: py4j, pyspark
    Running setup.py install for pyspark ... done
Successfully installed py4j-0.10.9 pyspark-3.0.2
```

Verify pyspark correctly installed:

```
% pyspark --version
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform
(file:/Library/Frameworks/Python.framework/Versions/3.9/lib/python3.9/site-packages/pyspark/jars/spark-unsafe_2.12-3.0.2.jar) to constructor java.nio.DirectByteBuffer(long,int)
WARNING: Please consider reporting this to the maintainers of org.apache.spark.unsafe.Platform
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.0.2
      /_/

Using Scala version 2.12.10, OpenJDK 64-Bit Server VM, 14.0.2
Branch HEAD
Compiled by user centos on 2021-02-16T04:53:13Z
Revision 648457905c4ea7d00e3d88048c63f360045f0714
Url https://gitbox.apache.org/repos/asf/spark.git
Type --help for more information.
```

Download and unpack Apache Spark from here.
(Note: This might be redundant to the above installation of pyspark. And suppose I unpacked it in my home directory.)

```
% cd ~
% tar tgz spark-3.0.2-bin-hadoop2.7.tgz
```

Create a symlink from /opt/spark to my unpacked spark directory:

```
% sudo ln -s ~/spark-3.0.2-bin-hadoop2.7 /opt/spark
```

Install findspark:

```
% pip3 install findspark
Collecting findspark
  Downloading findspark-1.4.2-py2.py3-none-any.whl (4.2 kB)
Installing collected packages: findspark
Successfully installed findspark-1.4.2
```

Run jupyter notebook and use MLlib with an example.

==

Running a local/standalone spark cluster (reference)

Step 1: Start a standalone cluster with a master

```
$ cd $HOME/spark-3.0.2-bin-hadoop2.7/sbin
$ ./start-all.sh
starting org.apache.spark.deploy.master.Master, logging to
/Users/jmonsod/spark-3.0.2-bin-hadoop2.7/logs/spark-jmonsod-org.apache.spark.deploy.master.M
aster-1-johns-mbp.lan.out
localhost: Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
Password:
localhost: starting org.apache.spark.deploy.worker.Worker, logging to
/Users/jmonsod/spark-3.0.2-bin-hadoop2.7/logs/spark-jmonsod-org.apache.spark.deploy.worker.W
orker-1-johns-mbp.lan.out

$ cat
/Users/jmonsod/spark-3.0.2-bin-hadoop2.7/logs/spark-jmonsod-org.apache.spark.deploy.worker.W
orker-1-johns-mbp.lan.out
Spark Command: /Library/Java/JavaVirtualMachines/jdk-14.0.2.jdk/Contents/Home/bin/java -cp
/Users/jmonsod/spark-3.0.2-bin-hadoop2.7/conf/:/Users/jmonsod/spark-3.0.2-bin-hadoop2.7/jars
/* -Xmx1g org.apache.spark.deploy.worker.Worker --webui-port 8081 spark://johns-mbp.lan:7077
========================================
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
21/03/20 01:25:38 INFO Worker: Started daemon with process name: 9373@johns-mbp.lan
21/03/20 01:25:38 INFO SignalUtils: Registered signal handler for TERM
21/03/20 01:25:38 INFO SignalUtils: Registered signal handler for HUP
21/03/20 01:25:38 INFO SignalUtils: Registered signal handler for INT
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform
(file:/Users/jmonsod/spark-3.0.2-bin-hadoop2.7/jars/spark-unsafe_2.12-3.0.2.jar) to
constructor java.nio.DirectByteBuffer(long,int)
WARNING: Please consider reporting this to the maintainers of
org.apache.spark.unsafe.Platform
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access
operations
WARNING: All illegal access operations will be denied in a future release
21/03/20 01:25:39 WARN NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
```

```
21/03/20 01:25:39 INFO SecurityManager: Changing view acls to: jmonsod
21/03/20 01:25:39 INFO SecurityManager: Changing modify acls to: jmonsod
21/03/20 01:25:39 INFO SecurityManager: Changing view acls groups to:
21/03/20 01:25:39 INFO SecurityManager: Changing modify acls groups to:
21/03/20 01:25:39 INFO SecurityManager: SecurityManager: authentication disabled; ui acls
disabled; users  with view permissions: Set(jmonsod); groups with view permissions: Set();
users  with modify permissions: Set(jmonsod); groups with modify permissions: Set()
21/03/20 01:25:39 INFO Utils: Successfully started service 'sparkWorker' on port 50591.
21/03/20 01:25:39 INFO Worker: Starting Spark worker 192.168.86.38:50591 with 8 cores, 15.0
GiB RAM
21/03/20 01:25:39 INFO Worker: Running Spark version 3.0.2
21/03/20 01:25:39 INFO Worker: Spark home: /Users/jmonsod/spark-3.0.2-bin-hadoop2.7
21/03/20 01:25:39 INFO ResourceUtils:
==============================================================
21/03/20 01:25:39 INFO ResourceUtils: Resources for spark.worker:

21/03/20 01:25:39 INFO ResourceUtils:
==============================================================
21/03/20 01:25:39 INFO Utils: Successfully started service 'WorkerUI' on port 8081.
21/03/20 01:25:39 INFO WorkerWebUI: Bound WorkerWebUI to 0.0.0.0, and started at
http://johns-mbp.lan:8081
21/03/20 01:25:39 INFO Worker: Connecting to master johns-mbp.lan:7077...
21/03/20 01:25:40 INFO TransportClientFactory: Successfully created connection to
johns-mbp.lan/192.168.86.38:7077 after 29 ms (0 ms spent in bootstraps)
21/03/20 01:25:40 INFO Worker: Successfully registered with master
spark://johns-mbp.lan:7077
```

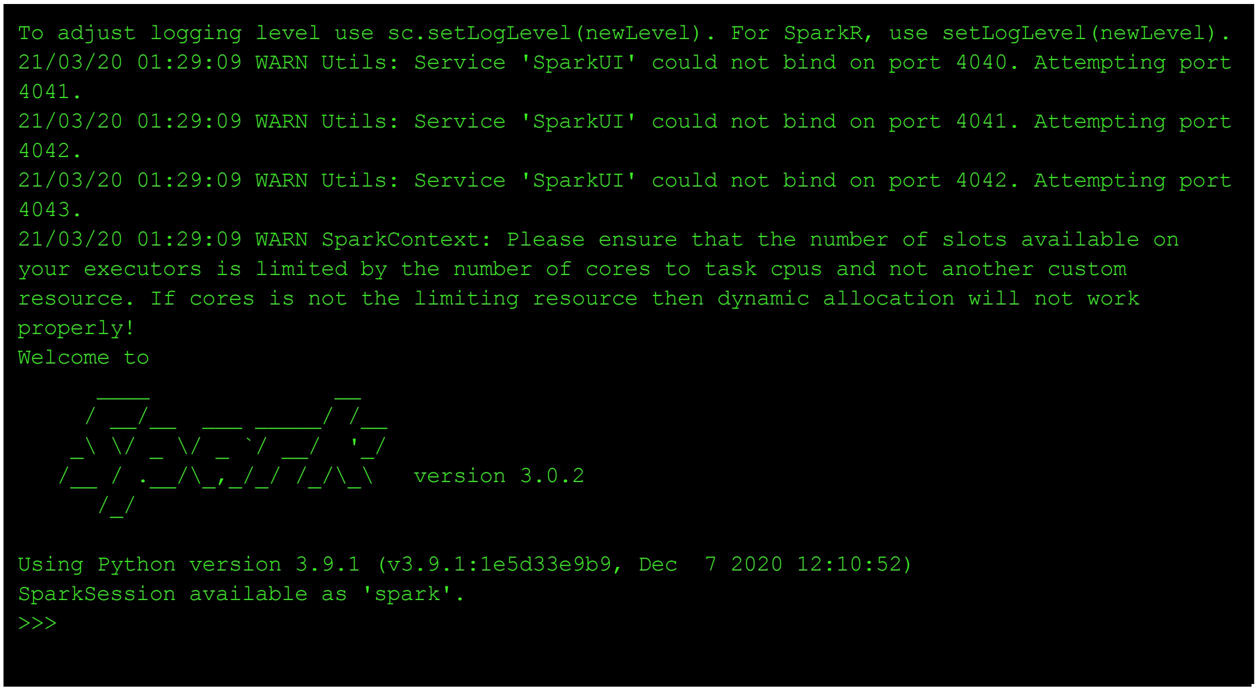
Note: if the command above complains like so:

localhost: ssh: connect to host localhost port 22: Connection refused

Then your laptop needs to enable remote login first (see this for a fix).

Step 2: Start an application that can attach to the cluster by specifying the master node. In this example, start a pyspark shell and have it point to the master:

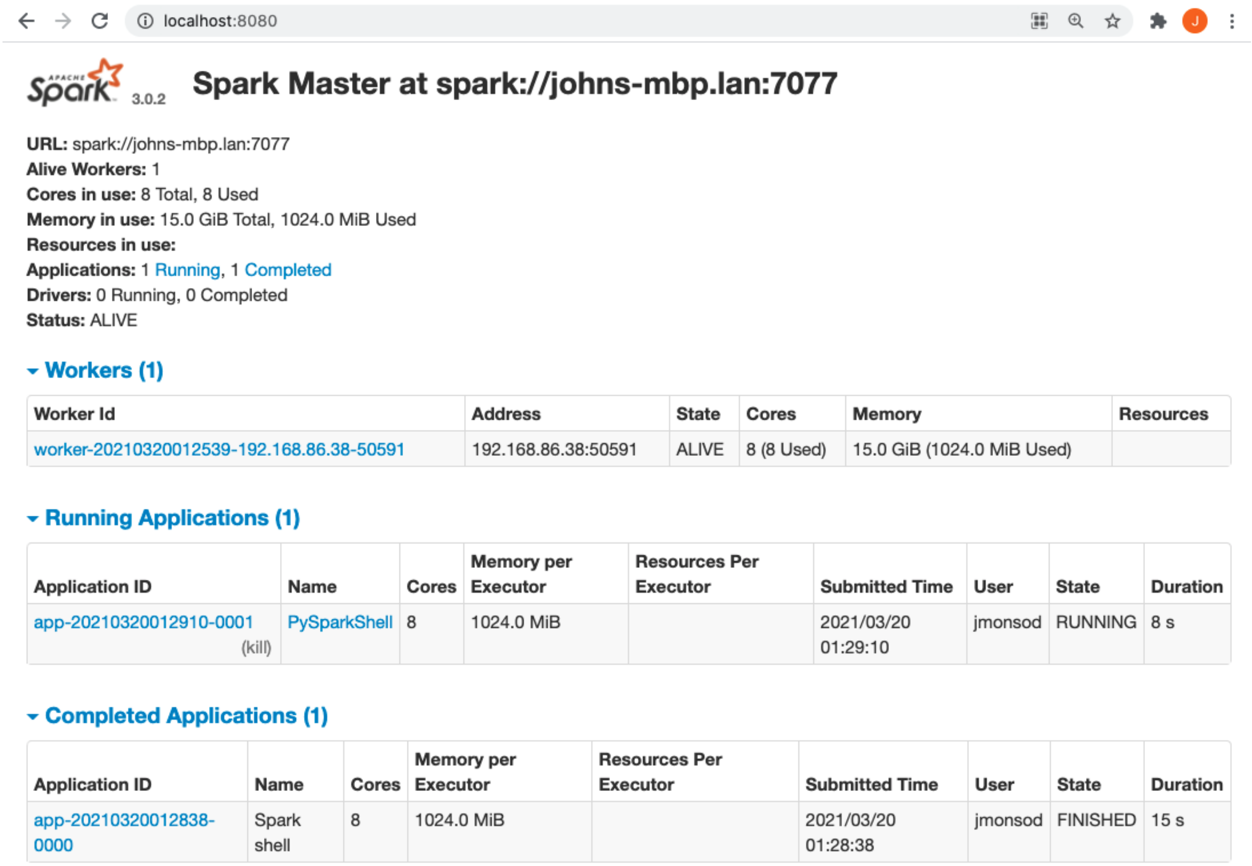
```
$ ../bin/pyspark --master spark://johns-mbp.lan:7077
Python 3.9.1 (v3.9.1:1e5d33e9b9, Dec  7 2020, 12:10:52)
[Clang 6.0 (clang-600.0.57)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform
(file:/Users/jmonsod/spark-3.0.2-bin-hadoop2.7/jars/spark-unsafe_2.12-3.0.2.jar) to
constructor java.nio.DirectByteBuffer(long,int)
WARNING: Please consider reporting this to the maintainers of
org.apache.spark.unsafe.Platform
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access
operations
WARNING: All illegal access operations will be denied in a future release
21/03/20 01:29:08 WARN NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
```



```
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
21/03/20 01:29:09 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port
4041.
21/03/20 01:29:09 WARN Utils: Service 'SparkUI' could not bind on port 4041. Attempting port
4042.
21/03/20 01:29:09 WARN Utils: Service 'SparkUI' could not bind on port 4042. Attempting port
4043.
21/03/20 01:29:09 WARN SparkContext: Please ensure that the number of slots available on
your executors is limited by the number of cores to task cpus and not another custom
resource. If cores is not the limiting resource then dynamic allocation will not work
properly!
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.0.2
      /_/

Using Python version 3.9.1 (v3.9.1:1e5d33e9b9, Dec  7 2020 12:10:52)
SparkSession available as 'spark'.
>>>
```

Check that everything is running as expected, browser: http://localhost:8080/, e.g.:



localhost:8080

3.0.2 Spark Master at spark://johns-mbp.lan:7077

URL: spark://johns-mbp.lan:7077
Alive Workers: 1
Cores in use: 8 Total, 8 Used
Memory in use: 15.0 GiB Total, 1024.0 MiB Used
Resources in use:
Applications: 1 Running, 1 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (1)

Worker Id	Address	State	Cores	Memory	Resources
worker-20210320012539-192.168.86.38-50591	192.168.86.38:50591	ALIVE	8 (8 Used)	15.0 GiB (1024.0 MiB Used)	

Running Applications (1)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20210320012910-0001 (kill)	PySparkShell	8	1024.0 MiB		2021/03/20 01:29:10	jmonsod	RUNNING	8 s

Completed Applications (1)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20210320012838-0000	Spark shell	8	1024.0 MiB		2021/03/20 01:28:38	jmonsod	FINISHED	15 s

As shown above, the pyspark shell is running as an application that the spark cluster recognizes.

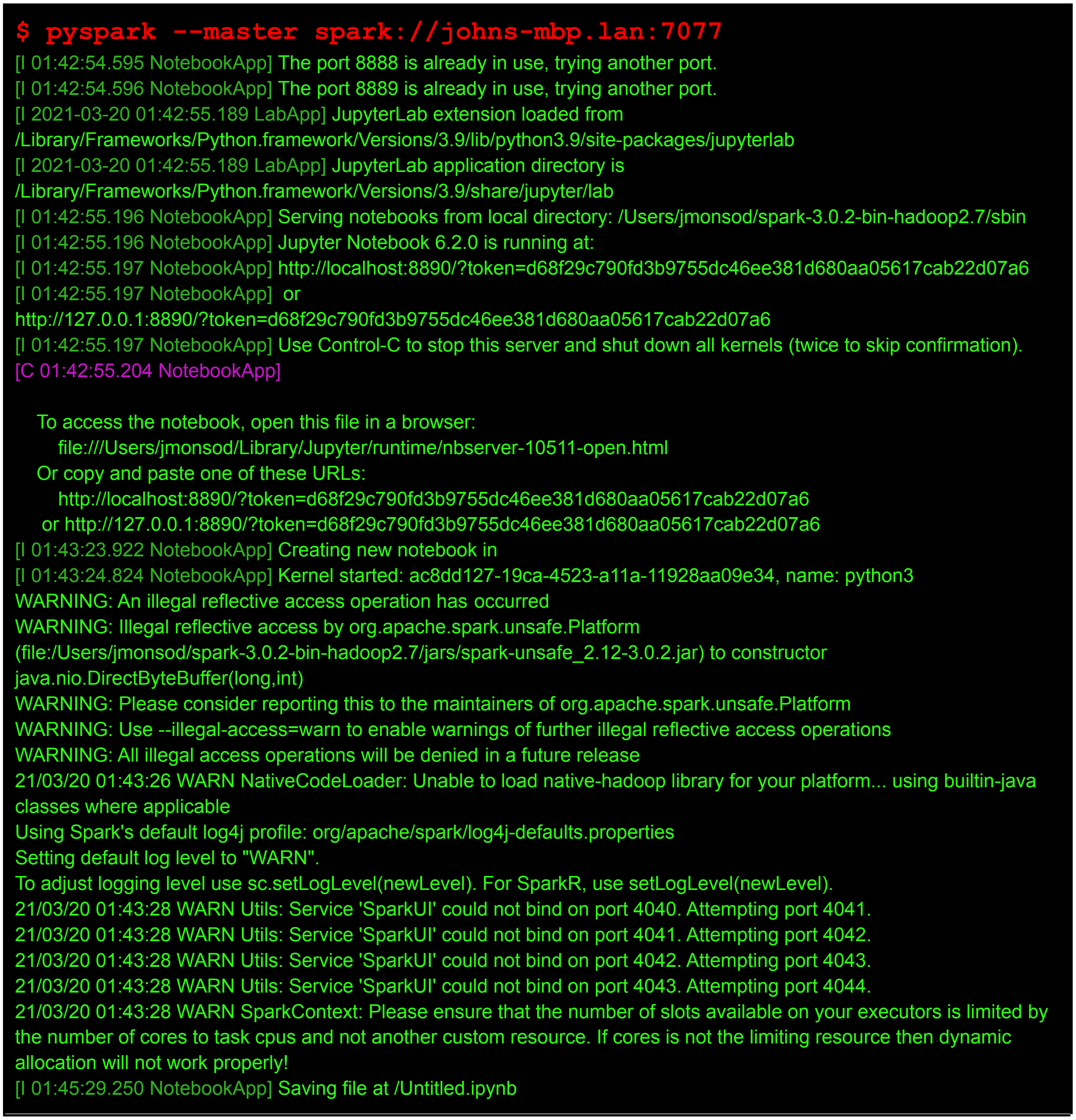
Step 3: Run a jupyter notebook as another application that is attached to the spark cluster.

If not yet set, ensure the following is in your env variables (this is in my ~/.zshrc file for my shell):

```
export SPARK_HOME=$HOME/spark-3.0.2-bin-hadoop2.7
export PYTHONPATH=${SPARK_HOME}/python:$PYTHONPATH
export PYTHONPATH=${SPARK_HOME}/python/lib/py4j-0.10.9-src.zip:$PYTHONPATH
export PYSPARK_PYTHON=python3
export PYSPARK_DRIVER_PYTHON=jupyter
export PYSPARK_DRIVER_PYTHON_OPTS=notebook

PATH=$PATH:$SPARK_HOME/bin
```

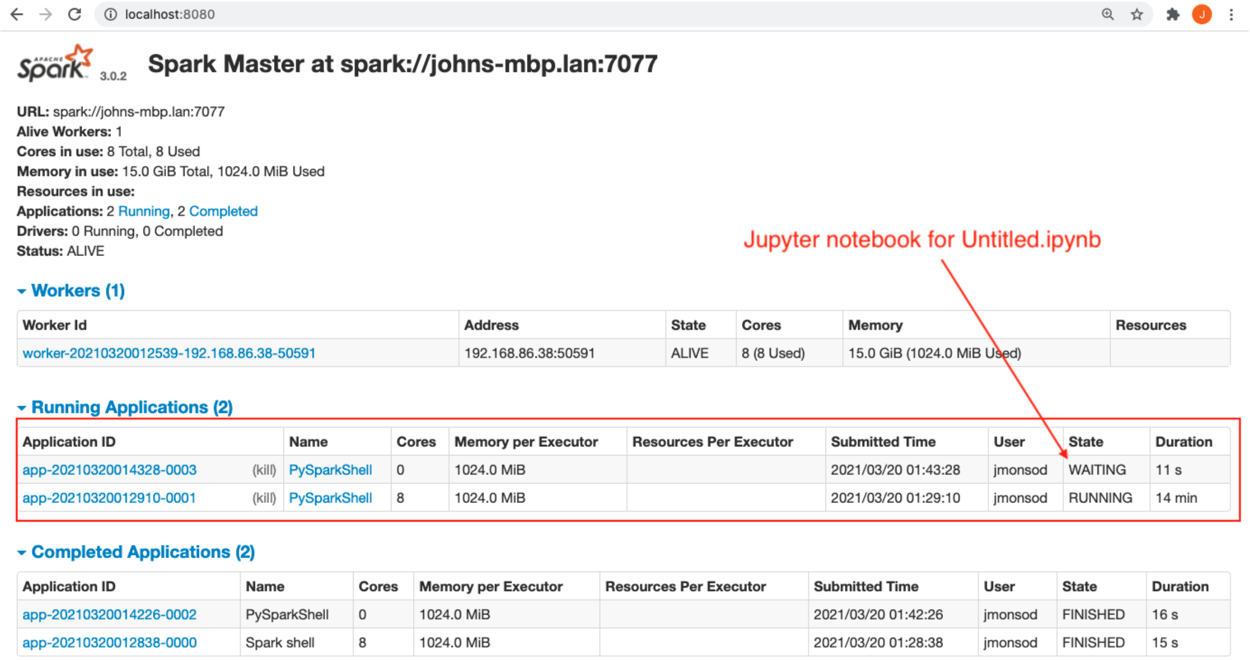
Starting pyspark with the ^^ env variables will automatically start jupyter instead of the pyspark shell on the command-line. As soon as a new notebook is created (e.g. Untitled.ipynb), it will be recognized by the spark cluster as another application.



```
$ pyspark --master spark://johns-mbp.lan:7077
[I 01:42:54.595 NotebookApp] The port 8888 is already in use, trying another port.
[I 01:42:54.596 NotebookApp] The port 8889 is already in use, trying another port.
[I 2021-03-20 01:42:55.189 LabApp] JupyterLab extension loaded from
/Library/Frameworks/Python.framework/Versions/3.9/lib/python3.9/site-packages/jupyterlab
[I 2021-03-20 01:42:55.189 LabApp] JupyterLab application directory is
/Library/Frameworks/Python.framework/Versions/3.9/share/jupyter/lab
[I 01:42:55.196 NotebookApp] Serving notebooks from local directory: /Users/jmonsod/spark-3.0.2-bin-hadoop2.7/sbin
[I 01:42:55.196 NotebookApp] Jupyter Notebook 6.2.0 is running at:
[I 01:42:55.197 NotebookApp] http://localhost:8890/?token=d68f29c790fd3b9755dc46ee381d680aa05617cab22d07a6
[I 01:42:55.197 NotebookApp]  or
http://127.0.0.1:8890/?token=d68f29c790fd3b9755dc46ee381d680aa05617cab22d07a6
[I 01:42:55.197 NotebookApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
[C 01:42:55.204 NotebookApp]

    To access the notebook, open this file in a browser:
        file:///Users/jmonsod/Library/Jupyter/runtime/nbserver-10511-open.html
    Or copy and paste one of these URLs:
        http://localhost:8890/?token=d68f29c790fd3b9755dc46ee381d680aa05617cab22d07a6
     or http://127.0.0.1:8890/?token=d68f29c790fd3b9755dc46ee381d680aa05617cab22d07a6
[I 01:43:23.922 NotebookApp] Creating new notebook in
[I 01:43:24.824 NotebookApp] Kernel started: ac8dd127-19ca-4523-a11a-11928aa09e34, name: python3
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform
(file:/Users/jmonsod/spark-3.0.2-bin-hadoop2.7/jars/spark-unsafe_2.12-3.0.2.jar) to constructor
java.nio.DirectByteBuffer(long,int)
WARNING: Please consider reporting this to the maintainers of org.apache.spark.unsafe.Platform
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
21/03/20 01:43:26 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
21/03/20 01:43:28 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
21/03/20 01:43:28 WARN Utils: Service 'SparkUI' could not bind on port 4041. Attempting port 4042.
21/03/20 01:43:28 WARN Utils: Service 'SparkUI' could not bind on port 4042. Attempting port 4043.
21/03/20 01:43:28 WARN Utils: Service 'SparkUI' could not bind on port 4043. Attempting port 4044.
21/03/20 01:43:28 WARN SparkContext: Please ensure that the number of slots available on your executors is limited by
the number of cores to task cpus and not another custom resource. If cores is not the limiting resource then dynamic
allocation will not work properly!
[I 01:45:29.250 NotebookApp] Saving file at /Untitled.ipynb
```

Verify it is recognized in the spark cluster:



Spark Master at spark://johns-mbp.lan:7077

URL: spark://johns-mbp.lan:7077
Alive Workers: 1
Cores in use: 8 Total, 8 Used
Memory in use: 15.0 GiB Total, 1024.0 MiB Used
Resources in use:
Applications: 2 Running, 2 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Jupyter notebook for Untitled.ipynb

Workers (1)

Worker Id	Address	State	Cores	Memory	Resources
worker-20210320012539-192.168.86.38-50591	192.168.86.38:50591	ALIVE	8 (8 Used)	15.0 GiB (1024.0 MiB Used)	

Running Applications (2)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20210320014328-0003 (kill)	PySparkShell	0	1024.0 MiB		2021/03/20 01:43:28	jmonsod	WAITING	11 s
app-20210320012910-0001 (kill)	PySparkShell	8	1024.0 MiB		2021/03/20 01:29:10	jmonsod	RUNNING	14 min

Completed Applications (2)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20210320014226-0002	PySparkShell	0	1024.0 MiB		2021/03/20 01:42:26	jmonsod	FINISHED	16 s
app-20210320012838-0000	Spark shell	8	1024.0 MiB		2021/03/20 01:28:38	jmonsod	FINISHED	15 s