Setting Up PySpark

CUNY The City College of New York

NYU Center for Urban Science + Progress

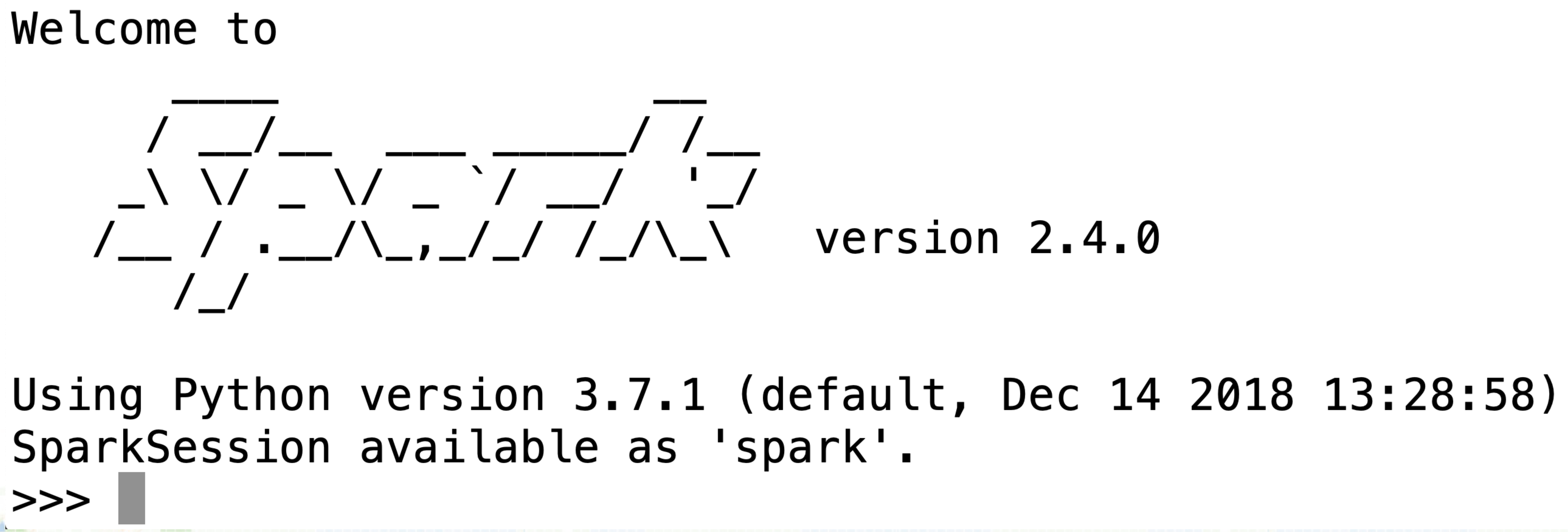
Install PySpark

- Through Anaconda, or run:

```
conda install pyspark
```

- Test pyspark

```
pyspark
```

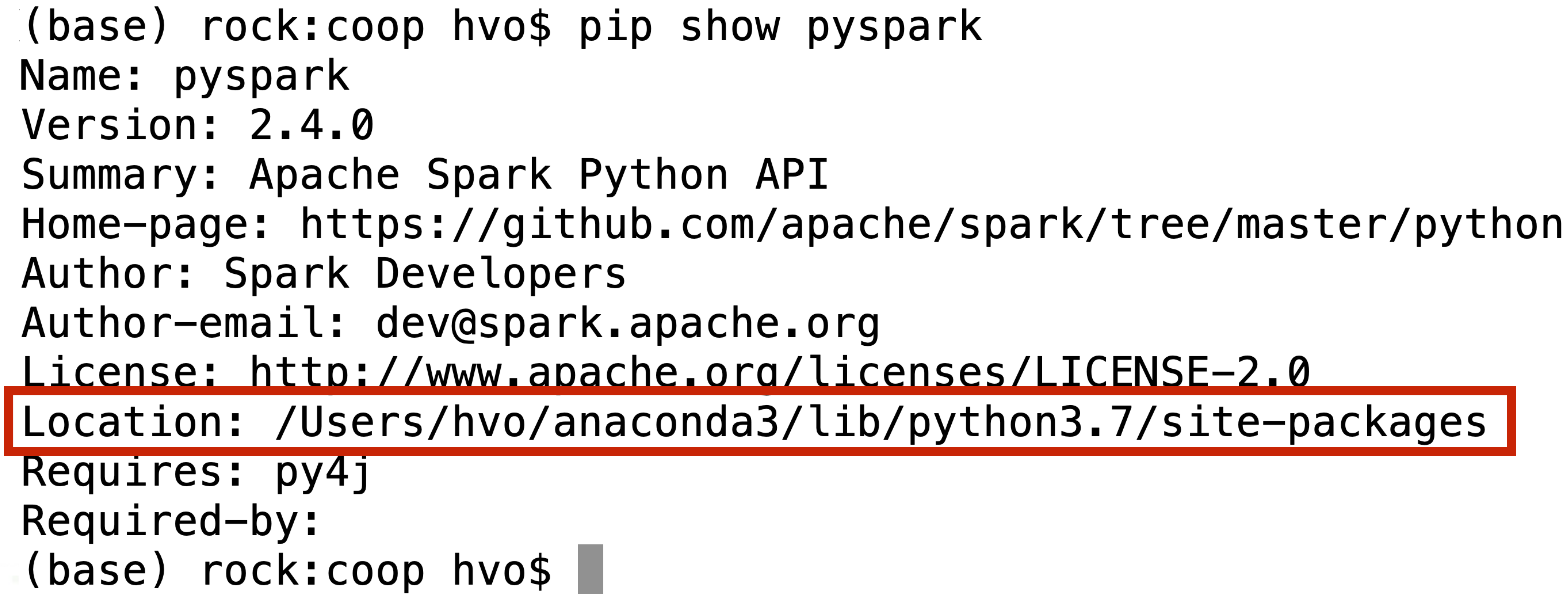


```
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.4.0
      /_/

Using Python version 3.7.1 (default, Dec 14 2018 13:28:58)
SparkSession available as 'spark'.
>>> 
```

Note the install location

```
pip show pyspark | grep Location
```



```
(base) rock:coop hvo$ pip show pyspark
Name: pyspark
Version: 2.4.0
Summary: Apache Spark Python API
Home-page: https://github.com/apache/spark/tree/master/python
Author: Spark Developers
Author-email: dev@spark.apache.org
License: http://www.apache.org/licenses/LICENSE-2.0
Location: /Users/hvo/anaconda3/lib/python3.7/site-packages
Requires: py4j
Required-by: 
(base) rock:coop hvo$ 
```

Mac/Linux — Environment Variables

```
export PYSPARK_DRIVER_PYTHON=`which jupyter`

export PYSPARK_DRIVER_PYTHON_OPTS=notebook

export SPARK_HOME=[PySpark Location]/pyspark
```

Windows — Environment Variables

```
set PYSPARK_PYTHON=jupyter

set PYSPARK_DRIVER_PYTHON=ipython

set PYSPARK_DRIVER_PYTHON_OPTS=notebook

set SPARK_HOME=[PySpark Location]/pyspark
```

Option 1 — Run Pyspark

- If you have exported the variables (in the last two slides), just run:

```
pyspark
```

- Else, in Bash (or Git-Bash) you can setup the variables right on the command line, run the below **all in one line** (separated by spaces):

```
PYSPARK_DRIVER_PYTHON=/Users/hvo/anaconda3/bin/jupyter
PYSPARK_DRIVER_PYTHON_OPTS=notebook
SPARK_HOME=/Users/hvo/anaconda3/lib/python3.7/site-packages/pyspark
/Users/hvo/anaconda3/bin/pyspark
```

Option 2 — Create a Jupyter Kernel

- Use the following repo to create your kernel

https://github.com/Anchormen/pyspark-jupyter-kernels

- In a Bash Terminal, run the following:

```
$ git clone https://github.com/Anchormen/pyspark-jupyter-kernels.git

$ cd pyspark-jupyter-kernels

$ ./pyspark_kernel.sh -t pyspark_kernel.template --spark_master local[*]
-d [Kernel Path] -k PySpark -e [VENV Path] --spark_home [PySpark Path]
```

options for these values are on the next slide

Kernel Configuration

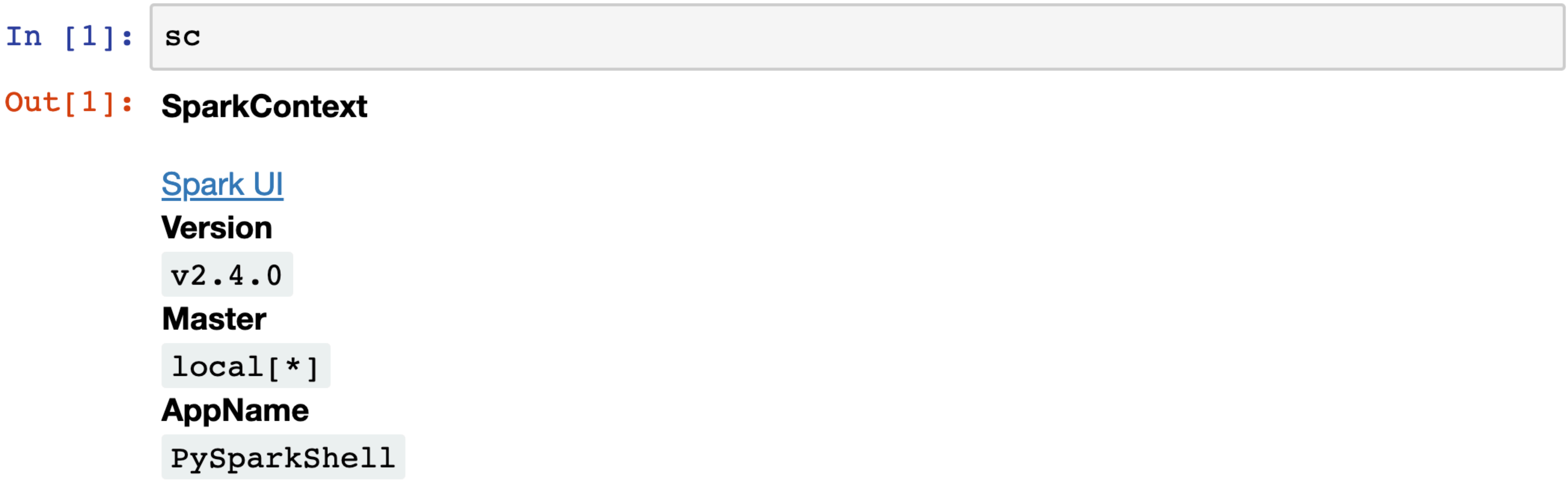
- Kernel Path, select one:
  - Linux: `~/.local/share/jupyter/kernels`
  - Mac: `~/Library/Jupyter/kernels`
  - Windows: `%APPDATA%\jupyter\kernels`
- VENV Path: your virtual environment path, find out by running:

  ```
  conda env list
  ```

  - and copy the path next to the *
- PySpark Path: the path that we have used before through "`pip show pyspark`"
- `local[*]`: this tells Spark to use all cores on your computer (*). To use a specific number, e.g. 4 cores, this should be set to `local[4]`.

Test Pyspark with Notebook

- Create a new notebook (use the kernel PySpark if you created one), and make sure the “**sc**” variable is valid:



```
In [1]: sc

Out[1]: SparkContext

        Spark UI
        Version
        v2.4.0
        Master
        local[*]
        AppName
        PySparkShell
```

Troubleshooting

- Try to install Java SE Development Kit

http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html

  - Restart your machine

- If it still doesn’t work, and you’re on Windows, try to run from source using the Instruction for Windows