Install Spark and Jupyter notebook.

DISCLAIMER

Last updated April 5, 2021

INTRODUCTION

The information provided by site https://tech2talk.github.io/vms-and-containers/ or https://github.com/tech2talk/vms-and-containers
is for informational purposes only. All information on the aforementioned sites and within source code or within documents contained
on them is provided "AS IS". I make no representation or warranty of any kind, express or implied, regarding the accuracy, adequacy,
validity, reliability, availability or completeness of any information contained or hosted within these sites.

Under no circumstance shall I have any liability to you or any user for any loss or damage of any kind incurred as a result of the use of
or reliance on any information provided on the aforementioned sites or the documents contained within them. Your use of the
aforementioned sites and your reliance on any information contained within documents including but not limited to source code is
solely at your own risk. This disclaimer template was created using Termly.

EXTERNAL LINKS DISCLAIMER

The documents stored or hosted on the aforementioned sites contain links to other websites or content belonging to or originating from third parties or links to websites and features in banners or other advertising. Such external links are not investigated, monitored, or checked for accuracy, adequacy, validity, reliability, availability or completeness by me.

I do not warrant, endorse, guarantee, or assume responsibility for the accuracy or reliability of any information offered by third-party websites mentioned in any documents or source code. I will not be a party to or in any way be responsible for monitoring any transaction between you and third-party providers of products or services.
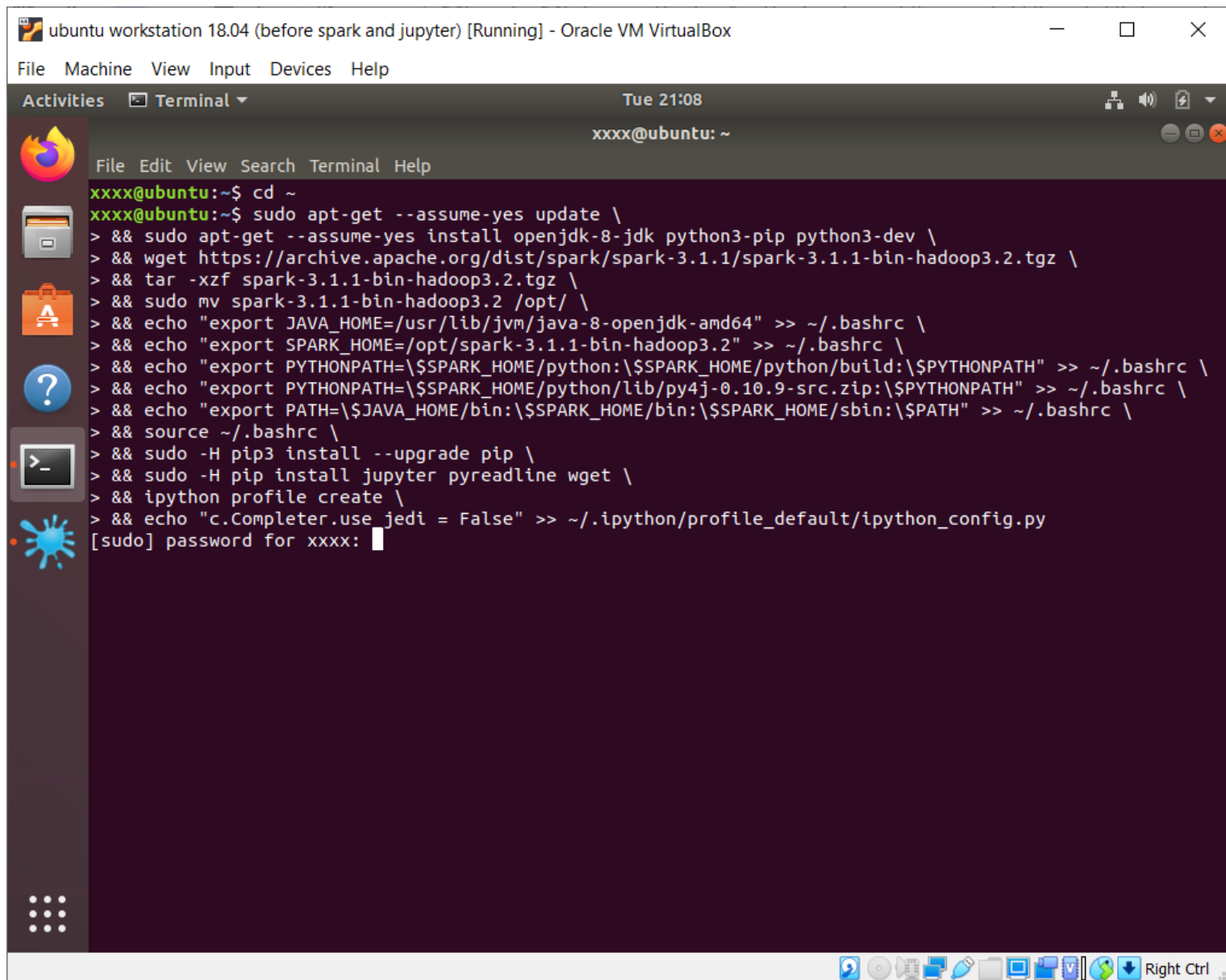
BY USING THESE WEBSITES (https://tech2talk.github.io/vms-and-containers/ or https://github.com/tech2talk/vms-and-containers **) OR THE INFORMATION CONTAINED ON THEM OR WITHIN ANY DOCUMENTS STORED IN THESE SITES INCLUDING BUT NOT LIMITED TO SOUCE CODE, YOU AGREE TO THE "DISCLAIMER" (INCLUDING "EXTERNAL LINKS DISCLAIMER") SPECIFIED IN ALL OF THE PREVIOUS PAGES AS WELL AS THE CURRENT PAGE OF THIS DOCUMENT.**

1. Start your Ubuntu virtual machine from VirtualBox. Login using the password and user name.

2. Visit this link and copy the entire command text from it

   https://github.com/tech2talk/vms-and-containers/blob/main/spark/spark-and-jupyter-setup-cmd

   Open a terminal and then paste the command you copied earlier to setup Spark & Jupyter notebook. Hit Enter to run these commands.

   Enter the password when prompted on the shell. It will take some time so let it run.

```
ubuntu workstation 18.04 (before spark and jupyter) [Running] - Oracle VM VirtualBox

File  Machine  View  Input  Devices  Help

Activities      Terminal                              Tue 21:08

                                  xxxx@ubuntu: ~

File  Edit  View  Search  Terminal  Help
xxxx@ubuntu:~$ cd ~
xxxx@ubuntu:~$ sudo apt-get --assume-yes update \
> && sudo apt-get --assume-yes install openjdk-8-jdk python3-pip python3-dev \
> && wget https://archive.apache.org/dist/spark/spark-3.1.1/spark-3.1.1-bin-hadoop3.2.tgz \
> && tar -xzf spark-3.1.1-bin-hadoop3.2.tgz \
> && sudo mv spark-3.1.1-bin-hadoop3.2 /opt/ \
> && echo "export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64" >> ~/.bashrc \
> && echo "export SPARK_HOME=/opt/spark-3.1.1-bin-hadoop3.2" >> ~/.bashrc \
> && echo "export PYTHONPATH=\$SPARK_HOME/python:\$SPARK_HOME/python/build:\$PYTHONPATH" >> ~/.bashrc \
> && echo "export PYTHONPATH=\$SPARK_HOME/python/lib/py4j-0.10.9-src.zip:\$PYTHONPATH" >> ~/.bashrc \
> && echo "export PATH=\$JAVA_HOME/bin:\$SPARK_HOME/bin:\$SPARK_HOME/sbin:\$PATH" >> ~/.bashrc \
> && source ~/.bashrc \
> && sudo -H pip3 install --upgrade pip \
> && sudo -H pip install jupyter pyreadline wget \
> && ipython profile create \
> && echo "c.Completer.use_jedi = False" >> ~/.ipython/profile_default/ipython_config.py
[sudo] password for xxxx:
```

3. Once the setup is complete, type the following command in the shell window which is open.
   mkdir -p ~/work/pyspark

4

The command above will create a directory called "work" in your home folder on Linux and then create a sub directory called "pyspark" within work directory.

4. Navigate to the newly created pyspark directory using the following command.
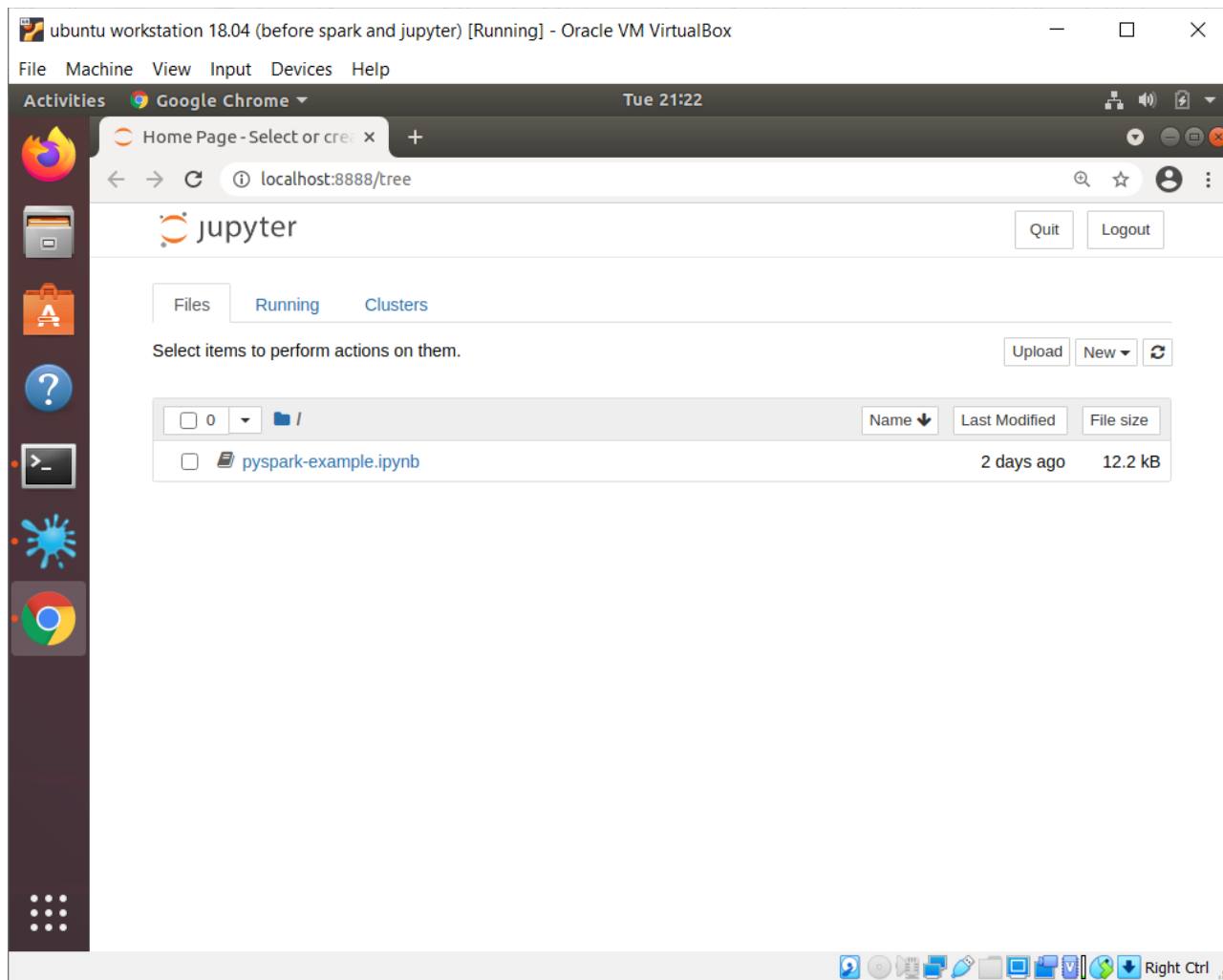
    cd ~/work/pyspark

5. Type the following command to download a sample exercise for Spark.

    wget https://raw.githubusercontent.com/tech2talk/vms-and-containers/main/spark/pyspark-example.ipynb
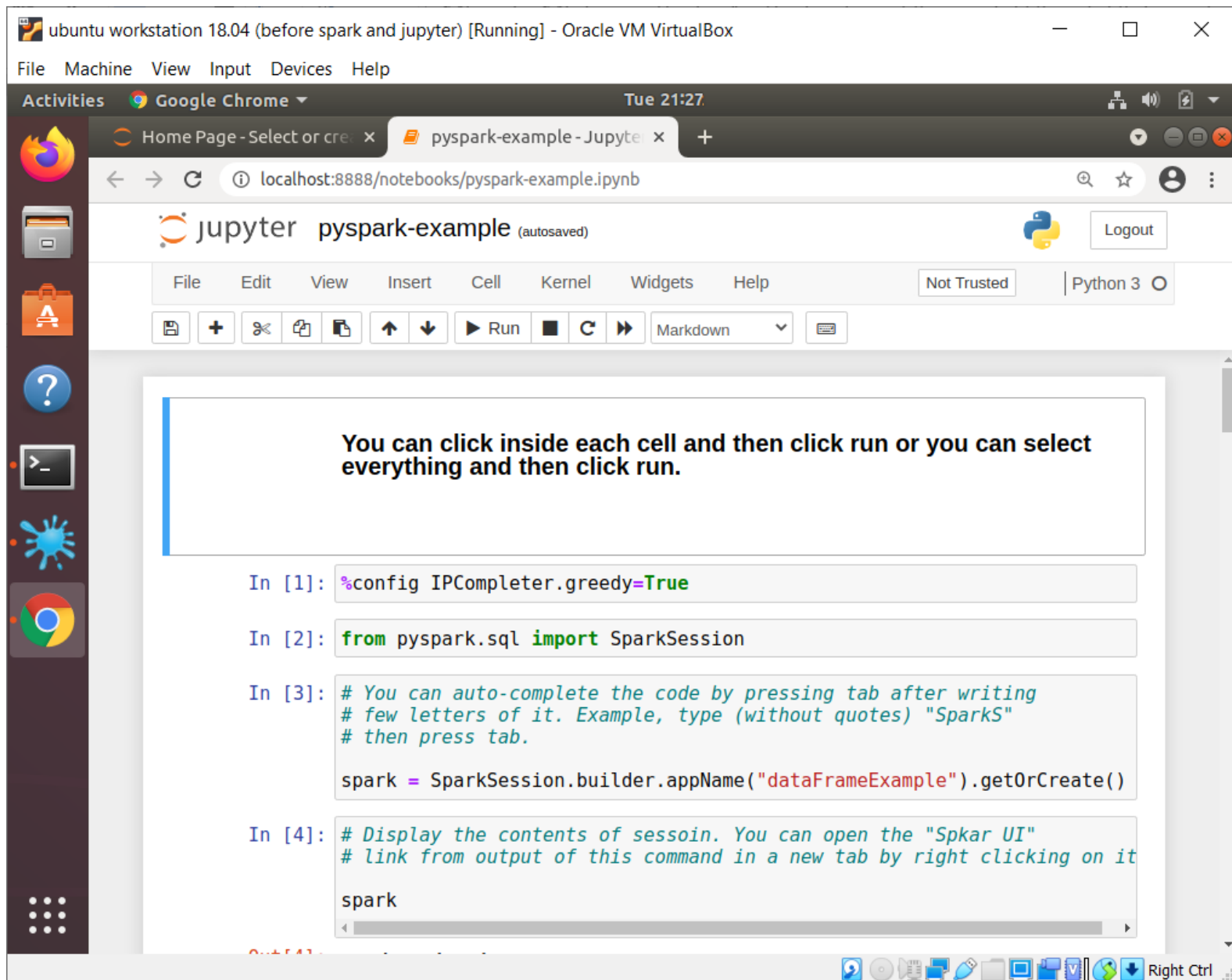
6. Type the following command to start Jupyter notebook.

    jupyter notebook

    It will start a new browser window like below. If it does not start then there might be an issue with the Step 2 setup.
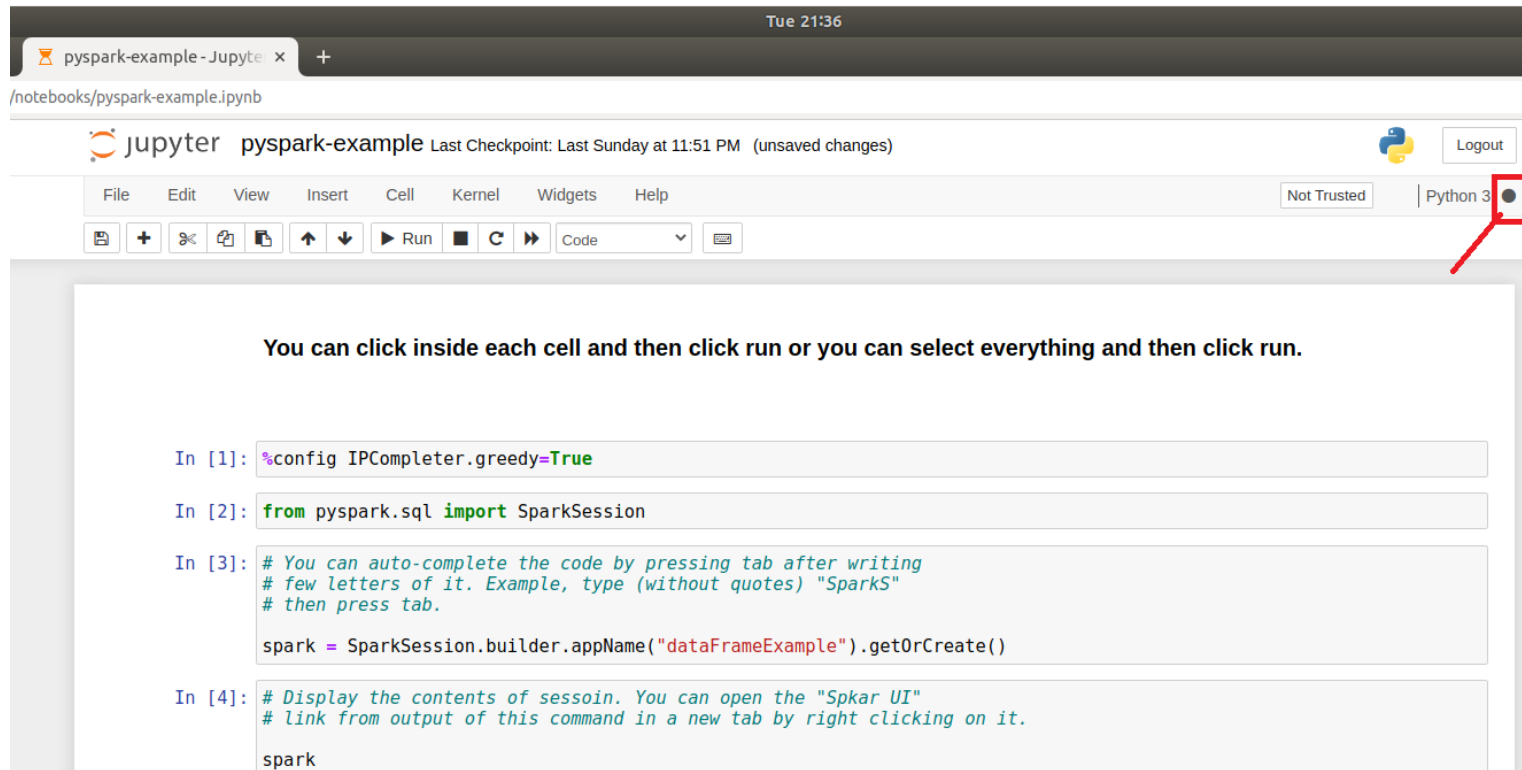
7. Click on the "pyspark-exmple.ipynb" to open the notebook. If you don't see this file then there might be an issue with Step 5.

It will open the notebook like below.

8. Follow the directions on the notebook to run the example code.

9. While the notebook is running command(s), the dot on the right upper corner will turn solid black.

Output of each command will be displayed below it.