Safari | O'Reilly
CLASS PREPARATION

Building Distributed Pipelines for Data Science Using Kafka, Spark, and Cassandra

Table of contents

Environment

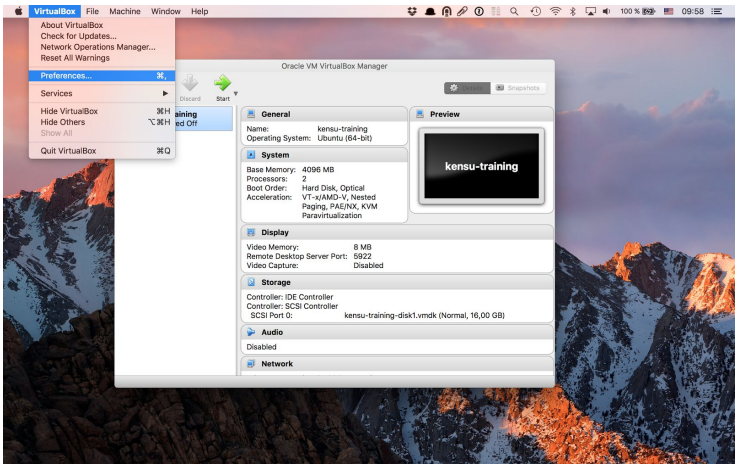
For the training, we'll be using as the simplest environment to run most of the pipeline.
This environment will be available from a single virtual machine image with Virtualbox.

Prerequisites

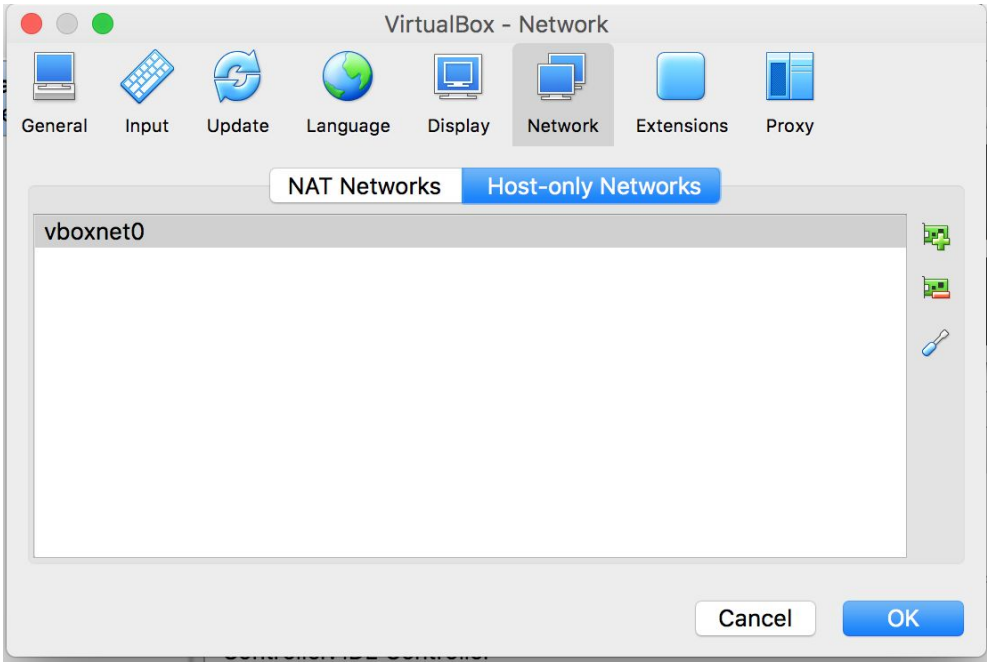
To use the training Virtual Machine, install Virtual Box (https://www.virtualbox.org) to your host.

Once it's installed, you need to make some customization before importing the training image. You first need to configure a "Host Only Network" to ensure you're able to communicate with the VM from your Host. If you have already one, ensure the DHCP server is enabled.

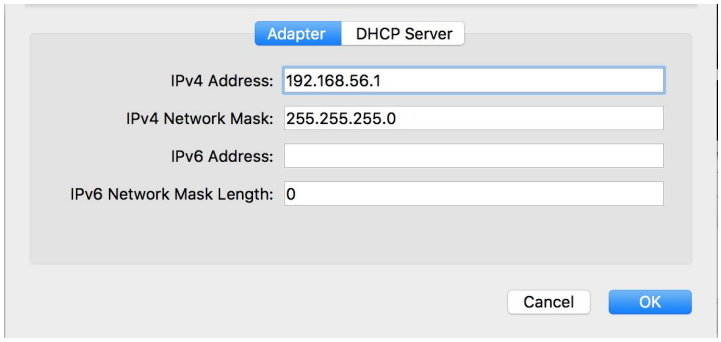
To do this, open VirtualBox Preferences:



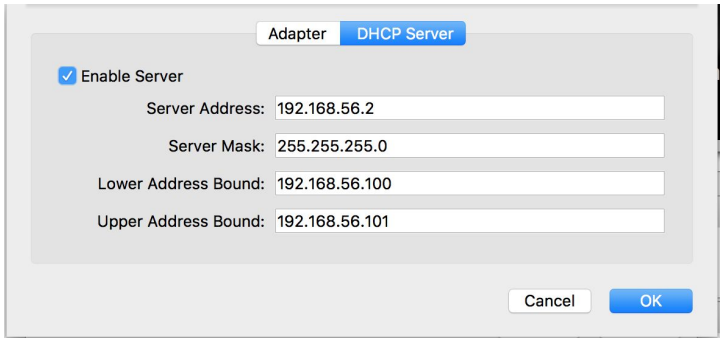
Click on the "Network" tab, then click on "Host-Only Networks" and finally click on the "+" button on the right.



The “vboxnet0” will be created. You also need to enable the DHCP server by selecting the “vboxnet0” and clicking on the screwdriver button.



Select the “DCHP Server” tab, ensure the checkbox “Enable Server” is ticked.



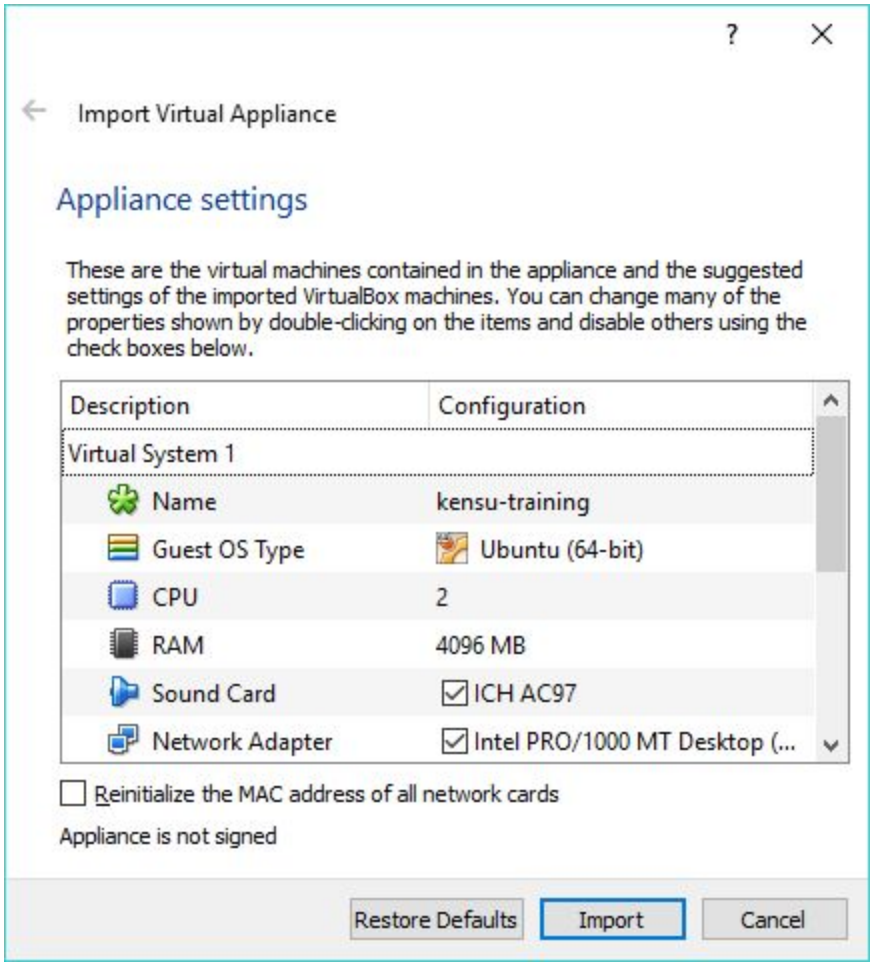
Note: if you’re running enterprise VPN like Cisco AnyConnect, please disconnect it before running the Virtual Machine, in some case that may conflict with the expected network.

Import the Virtual Machine

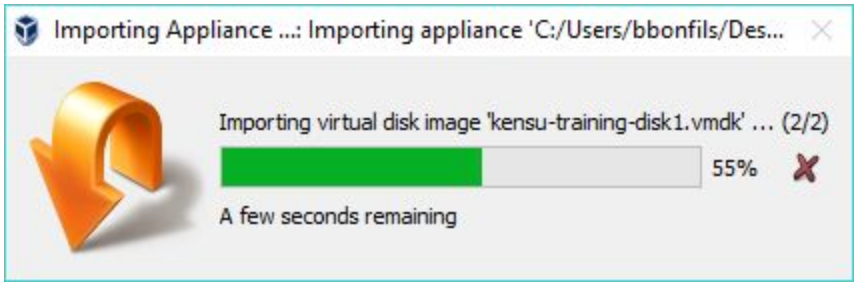
If vbox is running, shut down it.

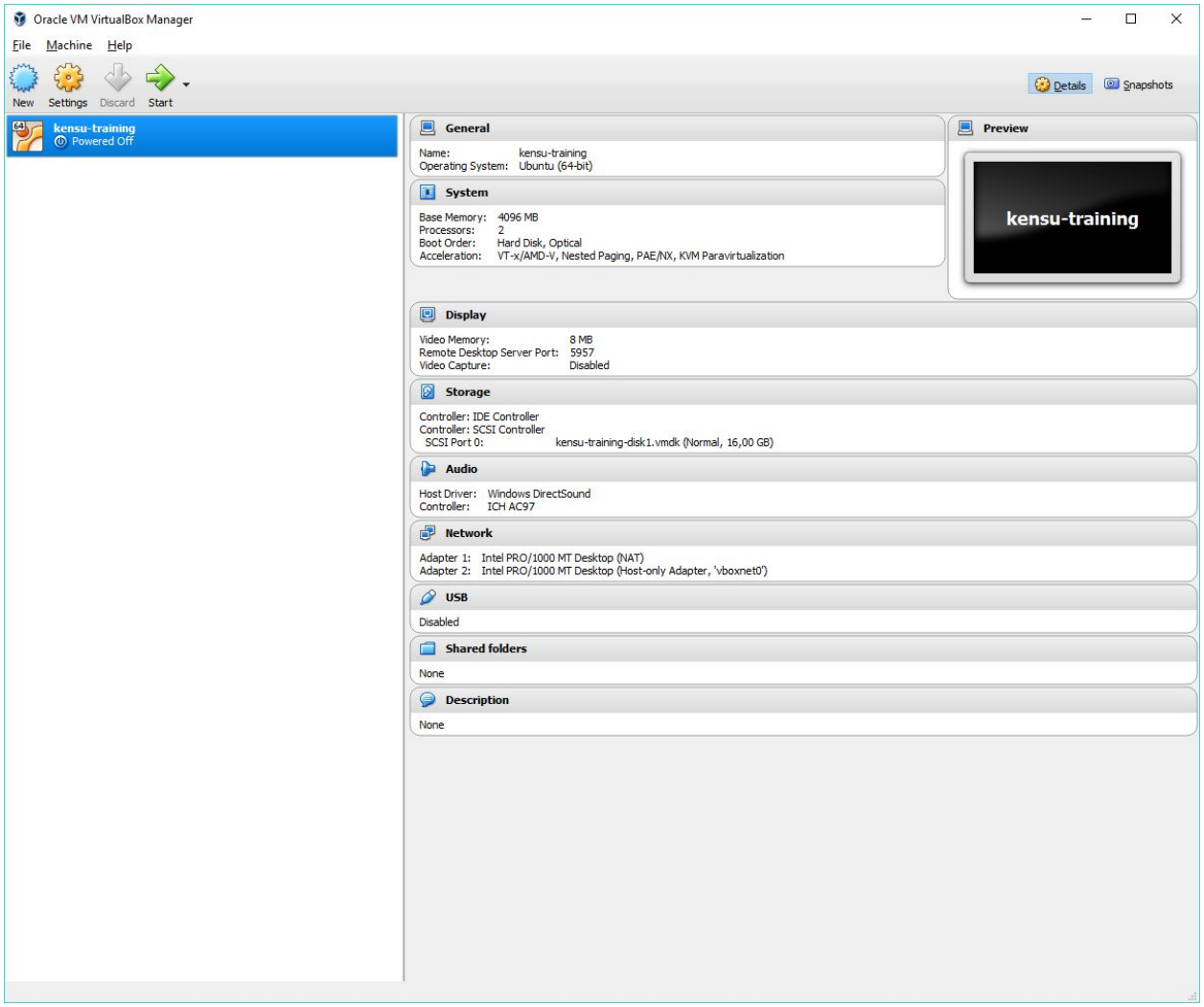
Download OVA file from http://kensuio-training.s3.amazonaws.com/kensu-training-00.ova then double-click on it.

For Chrome on Mac users: please check that the downloaded file have `ova` extension, depending your configuration, Chrome may have renamed it with `ovf` extension. In that case, please rename the file to get ova extension (see the chrome bug report).



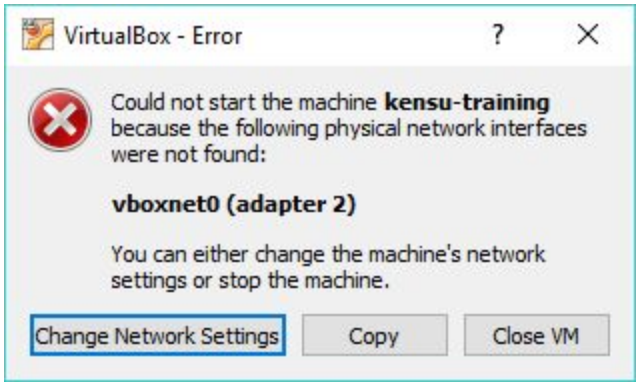
Click on import.



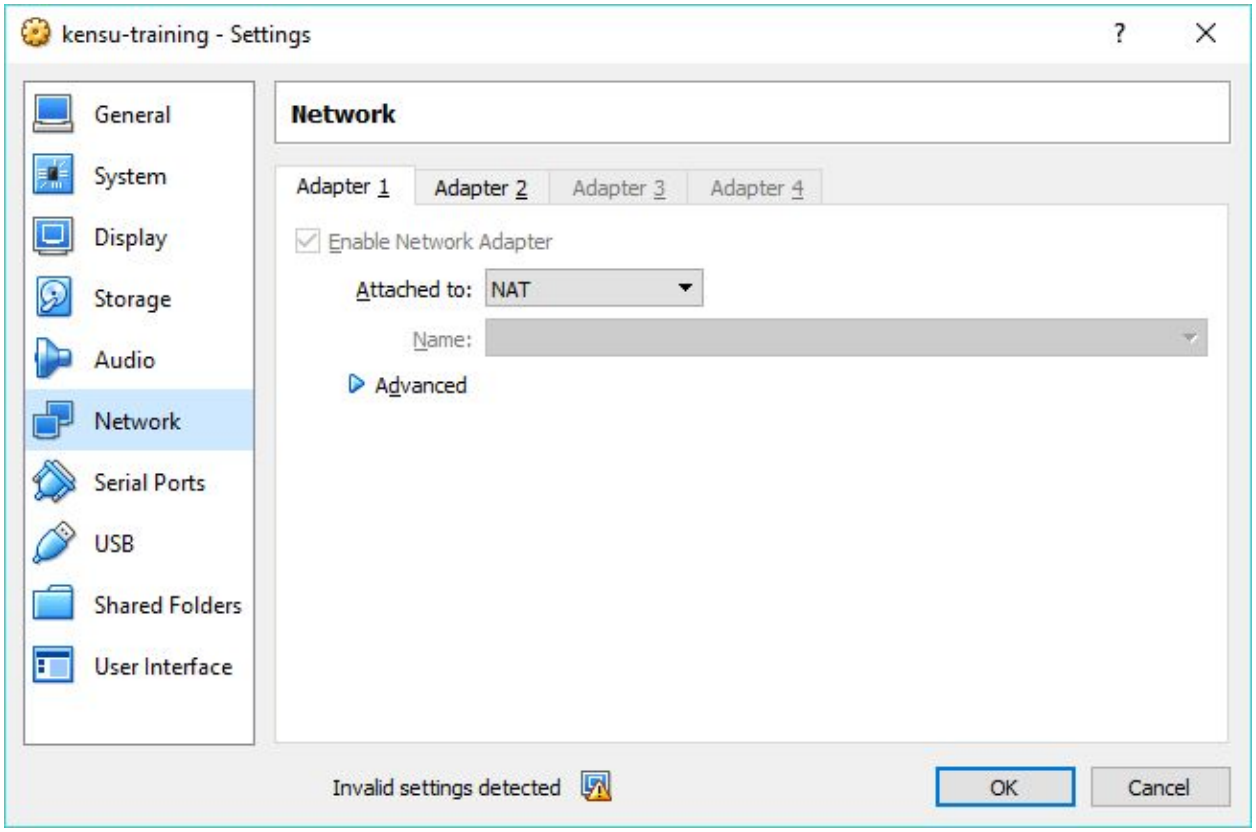


Select the VM and the click to “Start”

If you’re are running Windows, you’ll probably have this error screen:



Click on “Change Network Settings”



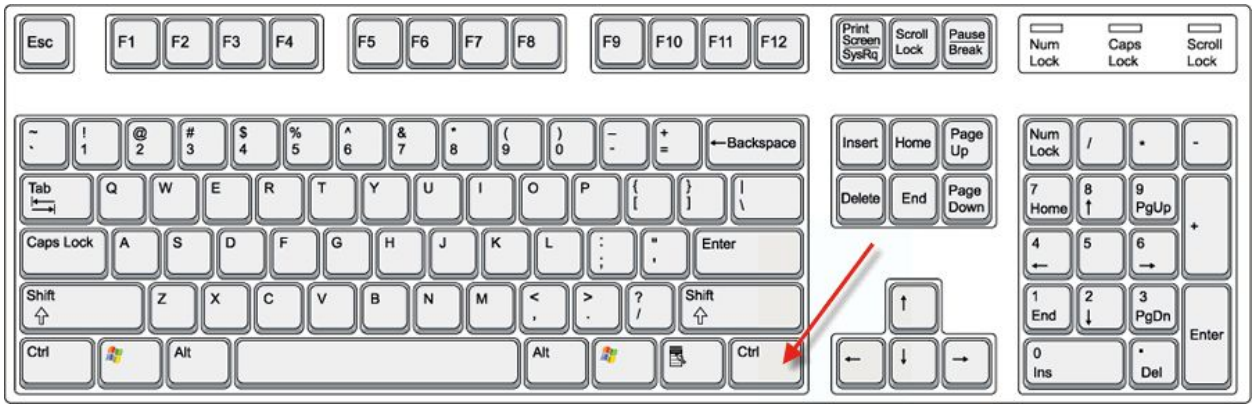
Then, click on “OK” button, no changes are required.

The virtual machine should be now running:

Now you can login with kensu as username, and kensu as password.

If you are stuck in Virtual Machine

REMEMBER THE DEFAULT KEY TO ESCAPE FROM THE VIRTUAL MACHINE IS **RIGHT CONTROL**

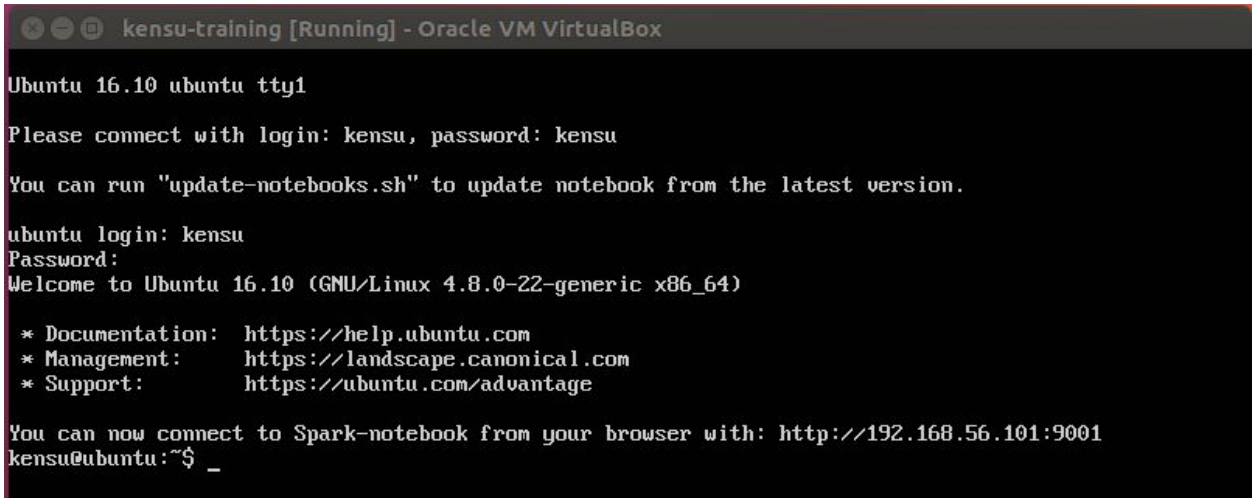


If you're stuck in the console, type "sudo halt -p"

Connecting to the Virtual Machine

After starting the Training VM in the step above, a shell will open, allowing you to connect to the VM.
Use the credentials:
User: `kensu`
Password: `kensu`

You will see a screen like this:



It is also recommended that you run "`update-notebooks.sh`" to get the latest updates to the notebooks used during the session.

Connect to Spark notebook

Using the output of the shell like illustrated above, copy the URL provided to access the notebook from your browser. (Chrome is strongly recommended, Firefox is OK. IE is not supported)
It's not absolutely necessary to login into the VM. It will often use the default IP address: 192.168.56.101, so after pressing "start" and waiting for the VM to boot, you could directly access the notebook at the URL: http://192.168.56.101:9000/
Note that this default IP Address might change. We strongly recommend login into the shell to get the exact address.

From the VM, we can always inspect the assigned local IP address issuing the following command:

```
kensu@ubuntu:~$ ip addr show dev eth1
3: eth1: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc pfifo_fast
state UP group default qlen 1000
    link/ether 08:00:27:a8:2e:32 brd ff:ff:ff:ff:ff:ff
    inet 192.168.56.101/24 brd 192.168.56.255 scope global eth1
       valid_lft forever preferred_lft forever
    inet6 fe80::a00:27ff:fea8:2e32/64 scope link
       valid_lft forever preferred_lft forever
```

In this example, the IP address is the text in bold.

Then, try to connect from your browser to http://192.168.56.101:9000/ (replace the IP address with the one you identified)

Stop the Virtual Machine

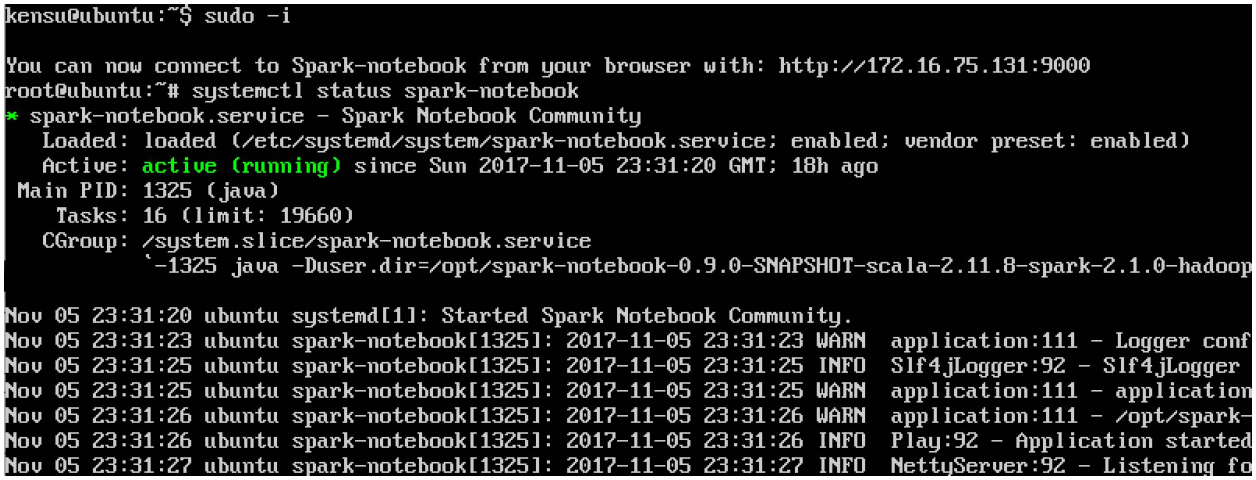
Connect to the Virtual Machine with login kensu, password kensu, and execute the following command:

```
kensu@ubuntu:~$: sudo halt -p
```

FAQ

SparkNotebook is not responding

1. test your network with “**ping 192.168.56.101**”
    a. no ping=network problem please check “prerequisites”
    b. login on console with kensu / kensu (or with ssh)
2. verify SparkNotebook status with “**systemctl status spark-notebook**”
    a. right status should look like this



    b. if you show an error/failed read the next step
3. become root with “**sudo -i**”
4. remove RUNNING_PID with “**rm /usr/share/spark-notebook/RUNNIG_PID**”
5. restart SparkNotebook with “**systemctl restart spark-notebook**”

How to update Notebooks ?

Run “update-notebooks.sh” to get the latest notebooks updates used during the session.

MacOSX Sierra network issue

there's a known bug on Sierra with VB for host-only network DHCP configuration.
https://apple.stackexchange.com/questions/278612/virtualbox-hangs-when-using-host-only-network