

CINF/CSCI 5931 Big Data Analytics

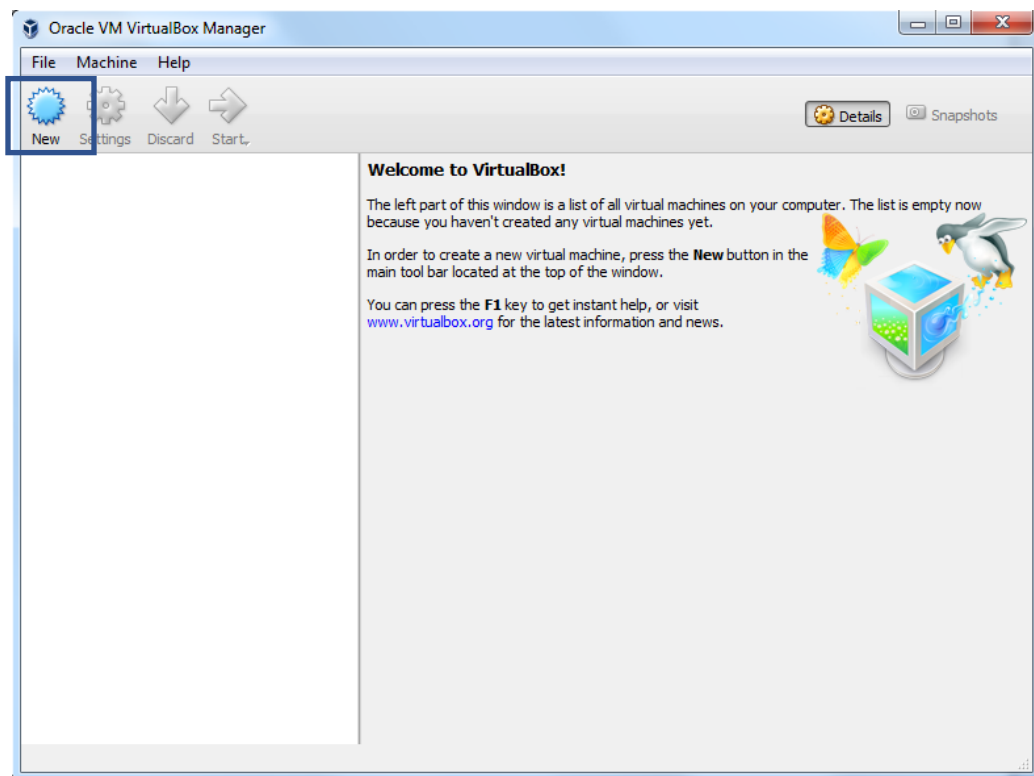
Lab Tutorial—Start Working with MLLib

Part I: Environment setup

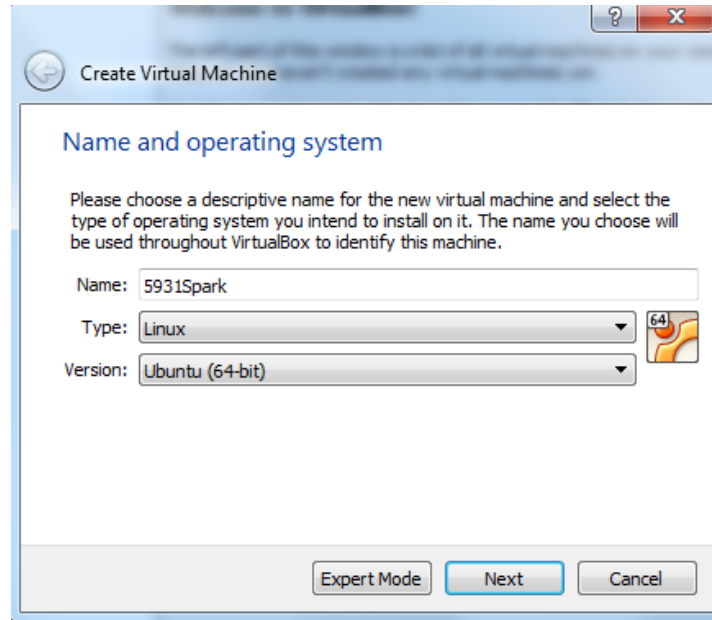
Last time, we introduced how to set up your Spark/Python environment using AWS. You can also set up environment on your local machine using virtual machine.

Step 1: Download and setup VirtualBox with Ubuntu

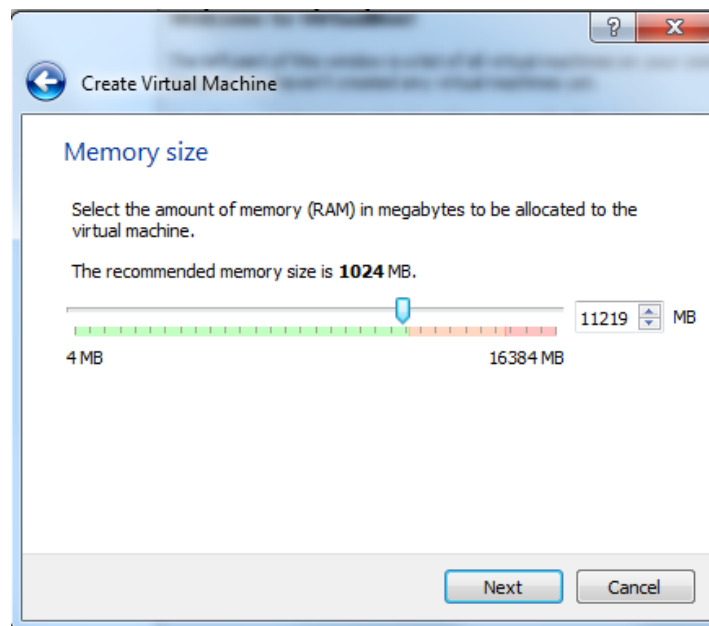
1. Download VirtualBox at <https://www.virtualbox.org/wiki/Downloads> (if you are not already running Ubuntu or other Linux system or already have other virtual machine).
2. Install VirtualBox and you will see the following.



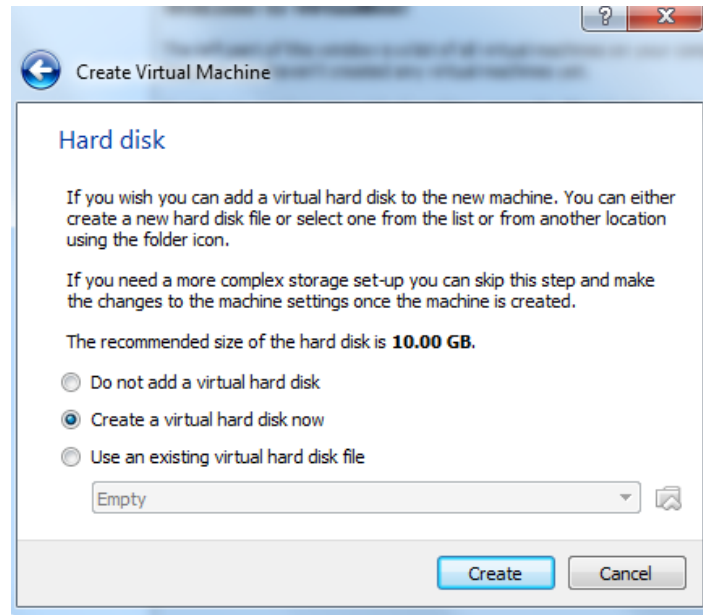
3. Go to <https://www.ubuntu.com/download/desktop> to download Ubuntu. The file is about 1.5GB so the download may take some time. Once the download is completed, you should see the Ubuntu iso file in your Downloads folder.
4. Open VirtualBox, create a new Virtual Machine as follows:
 - a. Click on New
 - b. Fill out the “Create Virtual Machine” pop-up window step by step.
 - i. For “Name and operating system”, name your machine, make “Type” be Linux, and choose your version based on your system



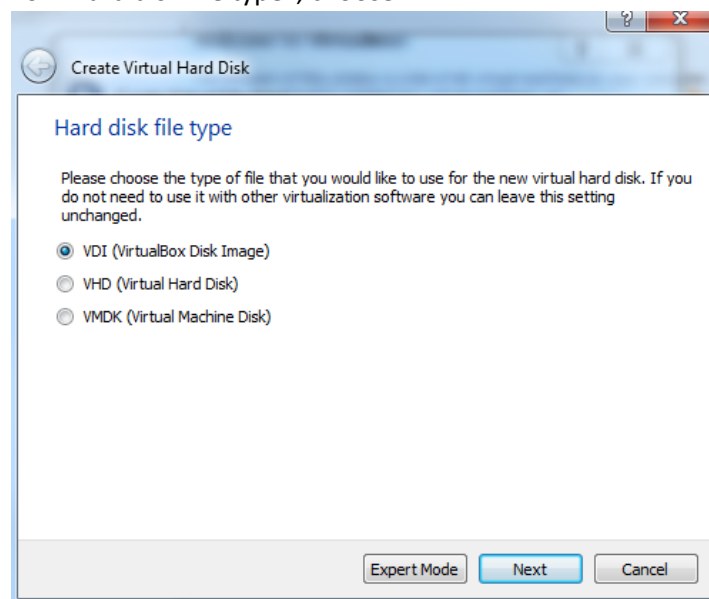
- ii. For “Memory size”, choose reasonable amount of memory based on what you have.



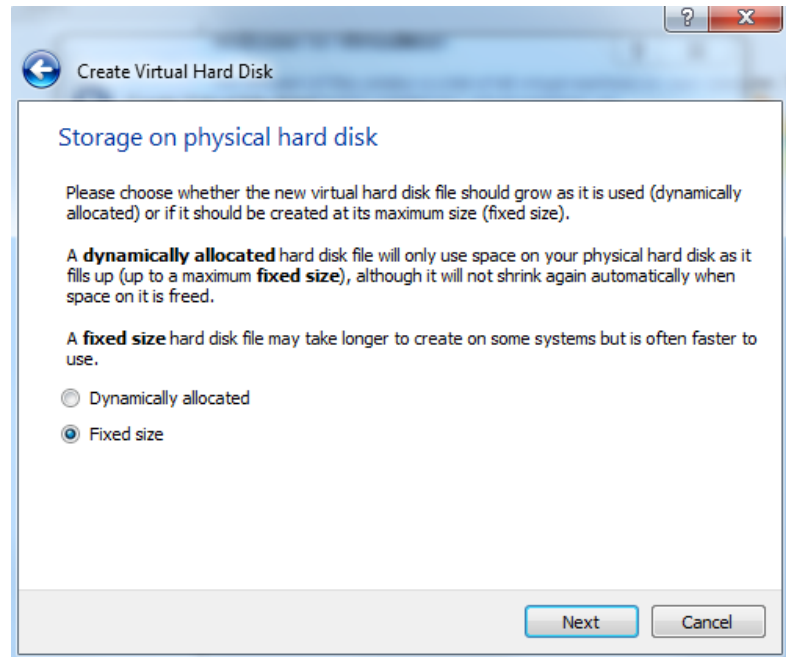
- iii. For “Hard disk”, choose “Create a virtual hard disk now”



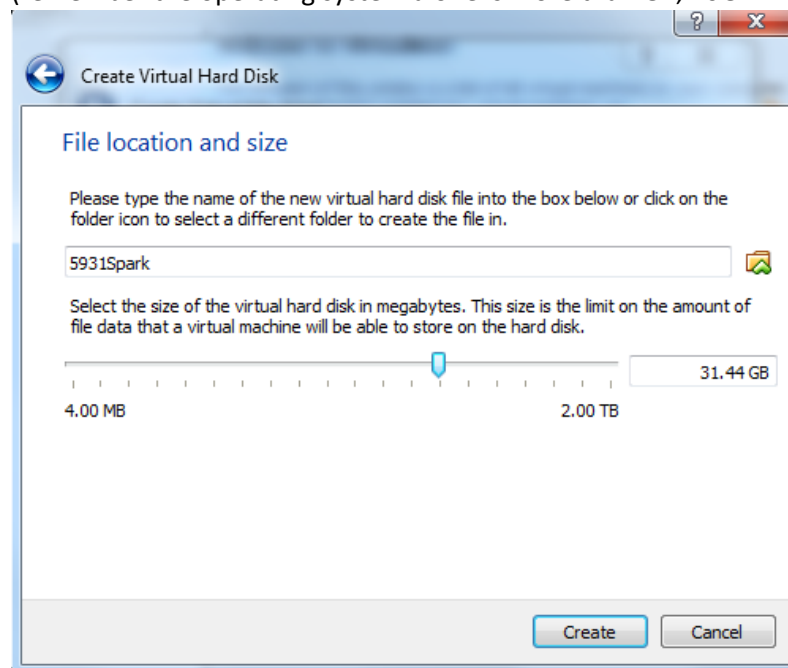
- iv. For “Hard disk file type”, choose “VDI”



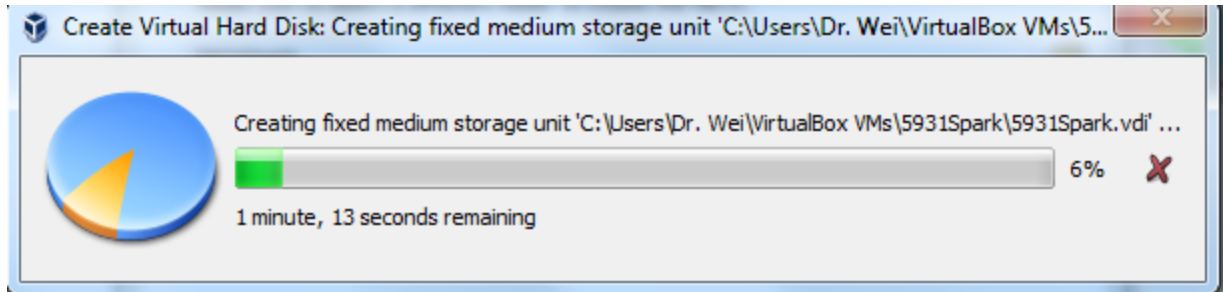
- v. For “Storage on physical hard disk”, choose “Fixed size”



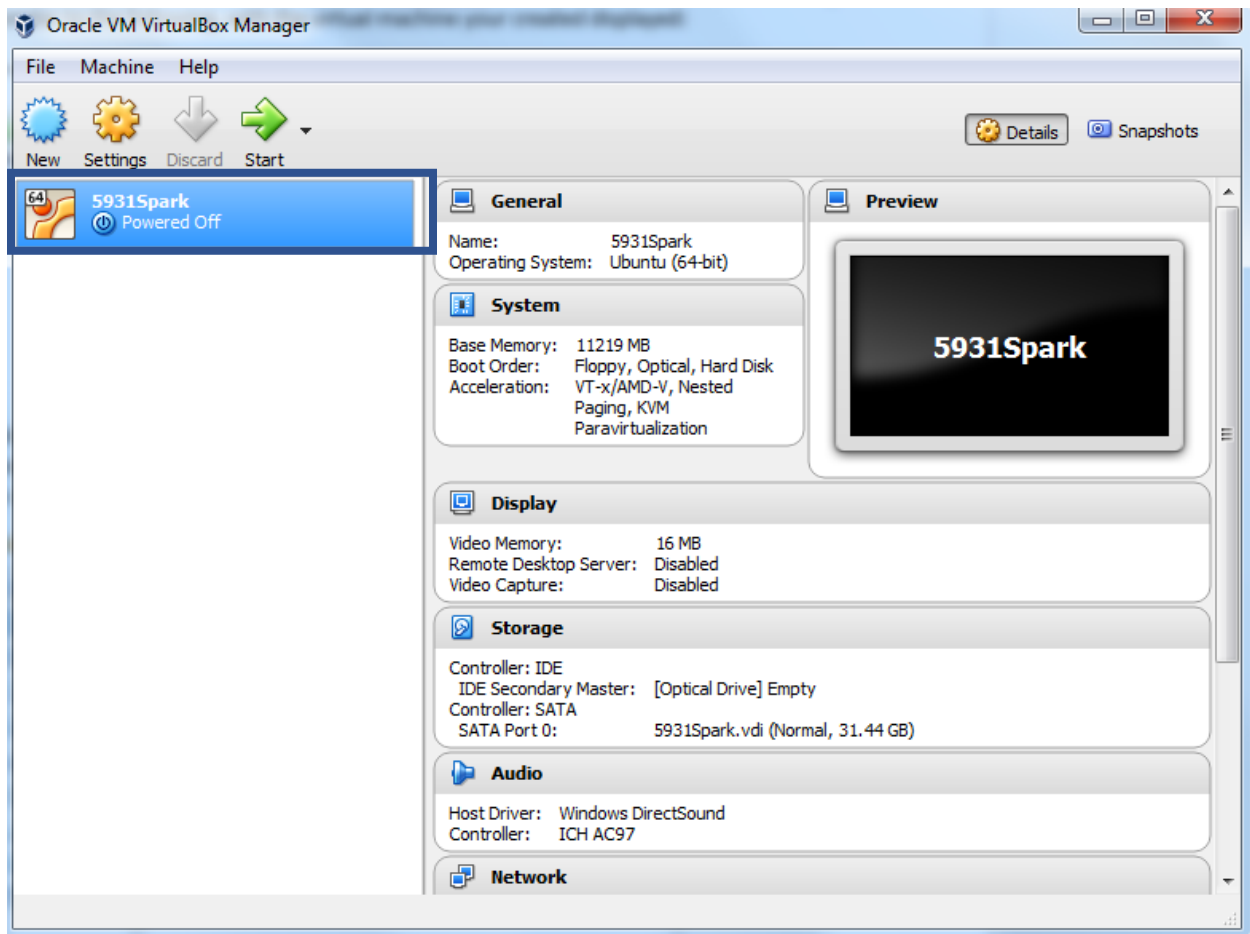
- vi. For “File location and size”, choose a reasonable size based on your system (remember the operating system alone is more than GB, 10GB minimum)



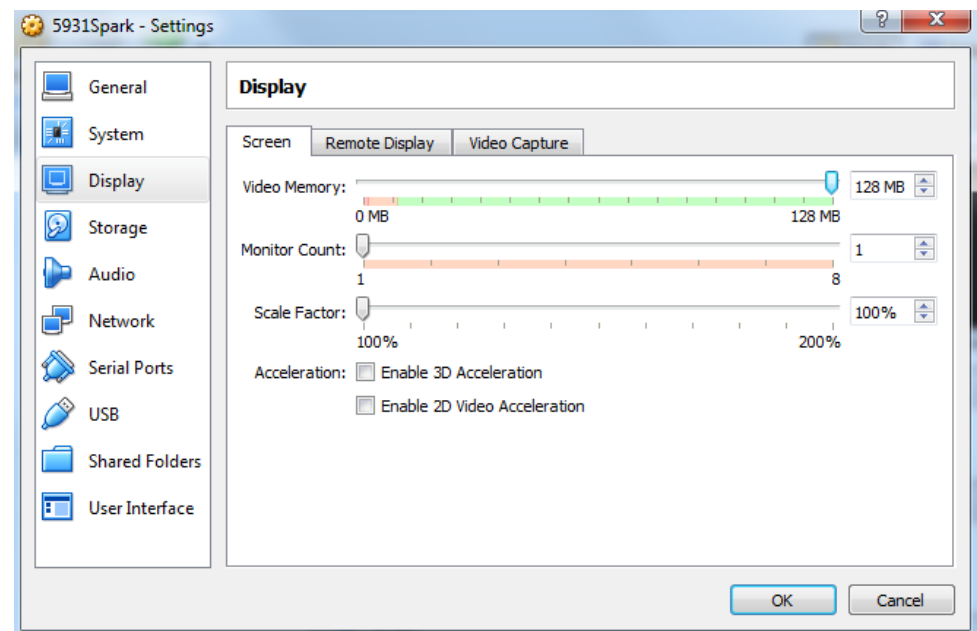
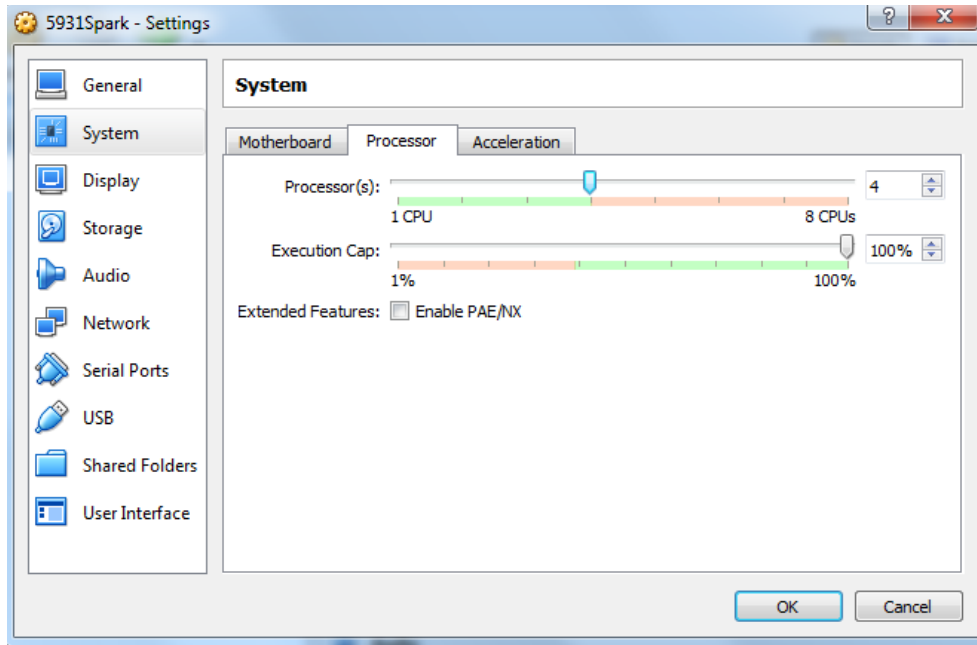
- vii. Click “Create”



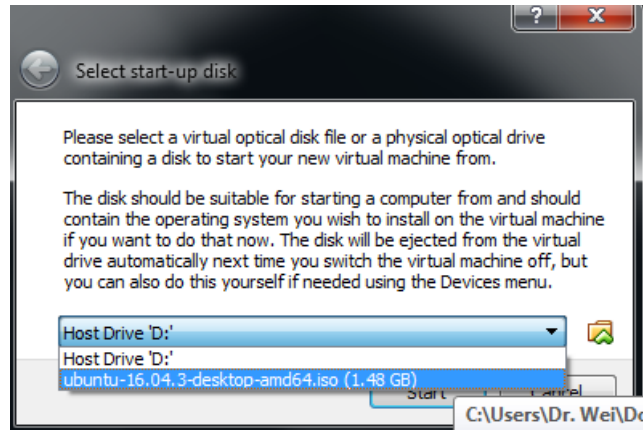
- c. Once the creation is done, through your VirtualBox manager you should see something similar to the following, with the virtual machine your created displayed:



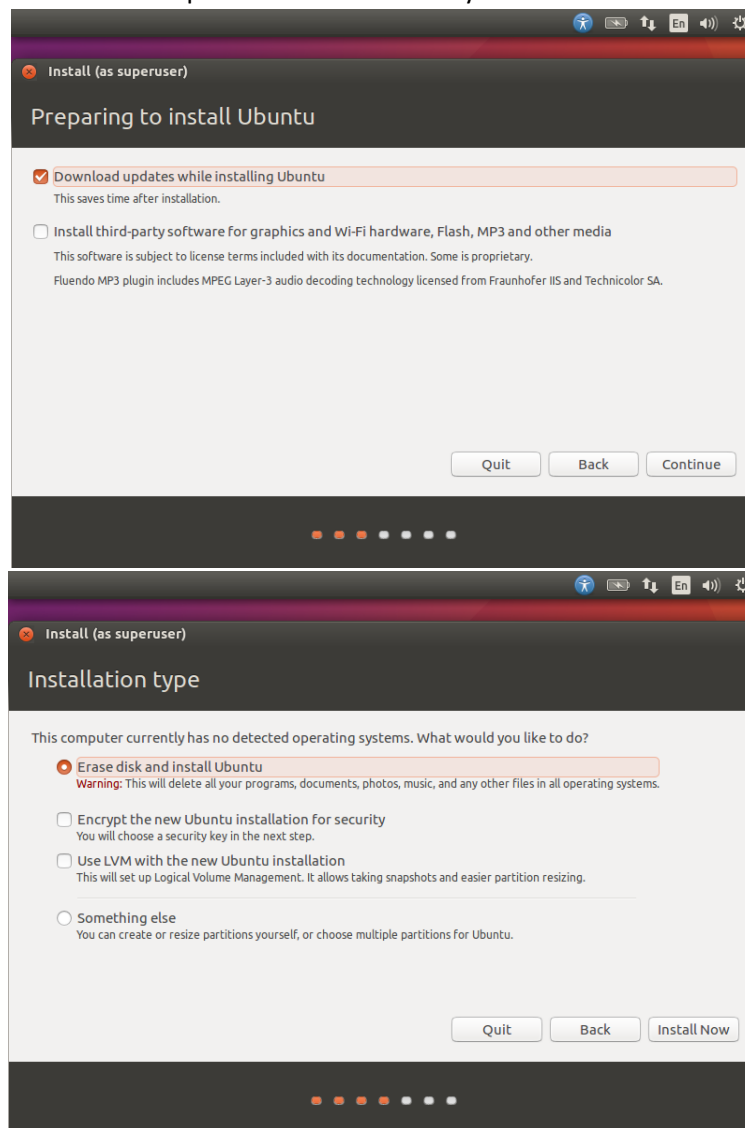
5. Right click on your virtual machine and choose "Settings". In the left panel, choose "System" and then choose the "Processor" tab, choose the number of processors (remain in the green zone). Also click on the "Screen" tab and choose proper "Video Memory". Click on "OK".

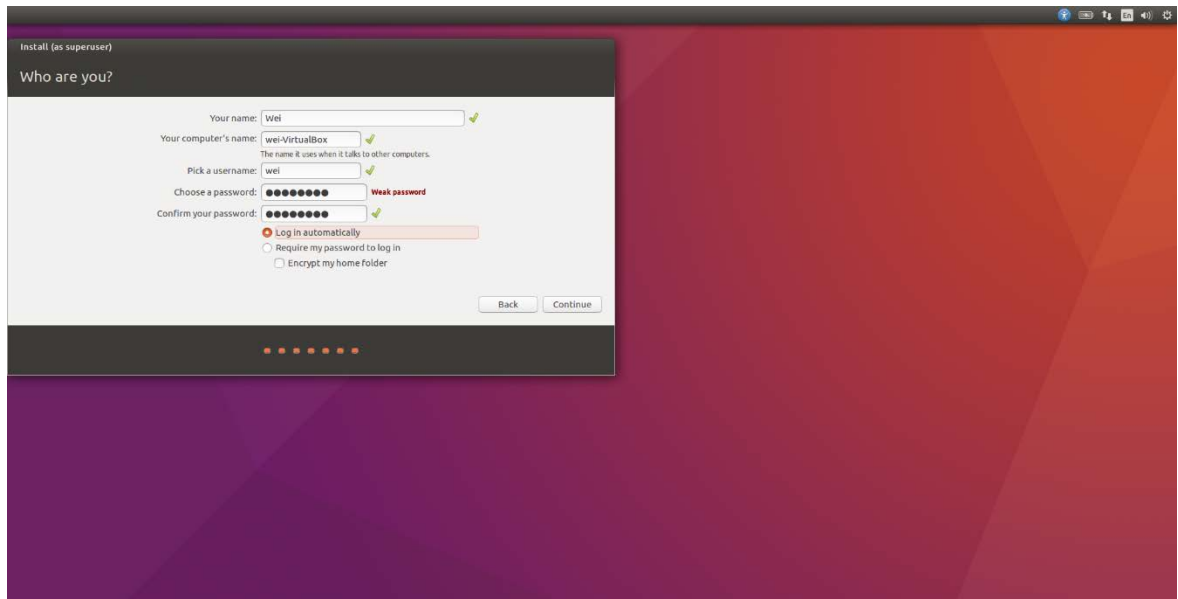


6. Double click on your virtual machine, and install Ubuntu on your virtual machine follow the step-by-step guide. Remember, you are installing Ubuntu on your virtual machine, not your actual machine.

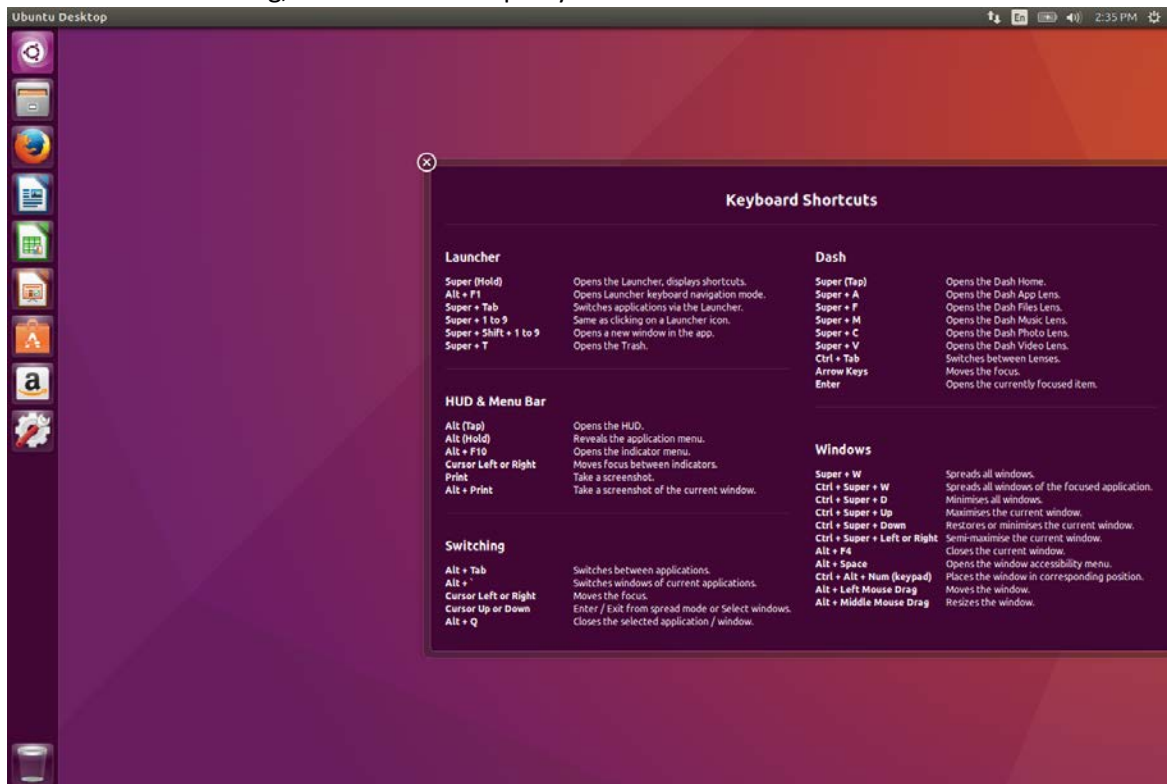


Then you need to follow the steps to install Ubuntu on your virtual machine:



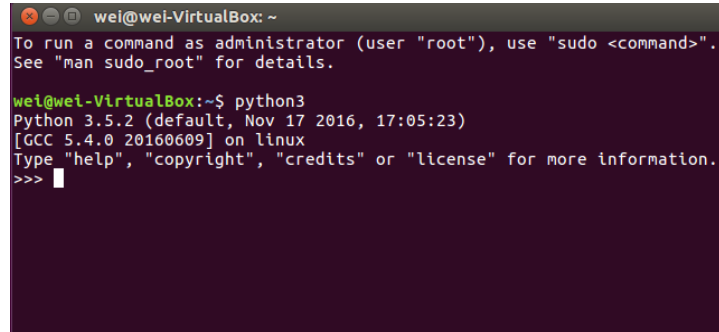


7. Once the installation is done, you restart your virtual machine and you should see something similar to the following, that is the desktop of your virtual machine:



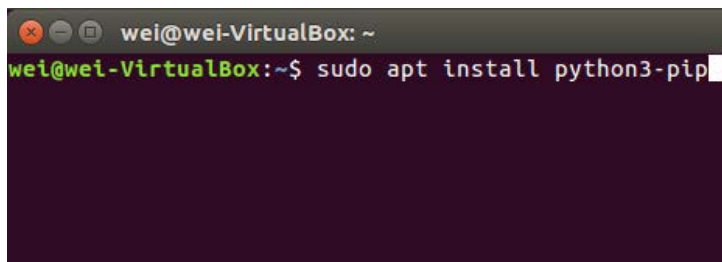
Step 2. Set up the Spark + Python environment on your virtual machine

1. Inside your virtual machine, open the terminal. Using command line to make sure you have python3.

A terminal window titled 'wei@wei-VirtualBox: ~' showing the execution of the 'python3' command. The output indicates that Python 3.5.2 is installed, with details about the default version, date, and GCC version. It also provides instructions on how to get help, copyright, credits, or license information.

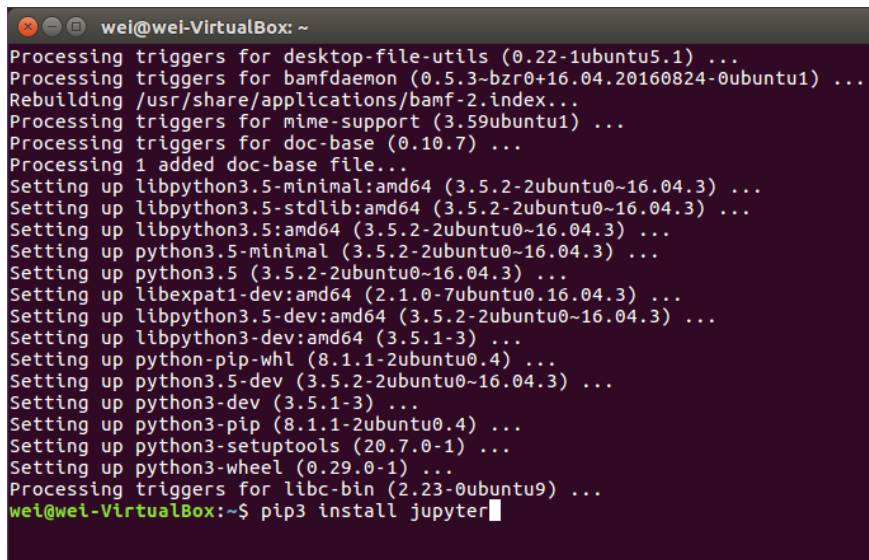
```
wei@wei-VirtualBox: ~  
To run a command as administrator (user "root"), use "sudo <command>".  
See "man sudo_root" for details.  
  
wei@wei-VirtualBox:~$ python3  
Python 3.5.2 (default, Nov 17 2016, 17:05:23)  
[GCC 5.4.0 20160609] on linux  
Type "help", "copyright", "credits" or "license" for more information.  
>>> 
```

2. Install jupyter notebook system, following the steps we did in setting up the environment on AWS.

A terminal window titled 'wei@wei-VirtualBox: ~' showing the execution of the command 'sudo apt install python3-pip'. The cursor is at the end of the command line, ready for execution.

```
wei@wei-VirtualBox:~$ sudo apt install python3-pip
```

(this installs pip3, we will use pip3 to install other python packages)

A terminal window titled 'wei@wei-VirtualBox: ~' showing the output of the 'sudo apt install python3-pip' command. The output lists various python3 packages being installed, including libpython3.5-minimal, libpython3.5-stdlib, libpython3.5, python3.5-minimal, python3.5, libexpat1-dev, libpython3.5-dev, libpython3-dev, python3-pip-whl, python3.5-dev, python3-dev, python3-pip, python3-setuptools, python3-wheel, and libc-bin. The command 'pip3 install jupyter' is also shown at the bottom.

```
wei@wei-VirtualBox: ~  
Processing triggers for desktop-file-utils (0.22-1ubuntu5.1) ...  
Processing triggers for bamfdaemon (0.5.3-bzr0+16.04.20160824-0ubuntu1) ...  
Rebuilding /usr/share/applications/bamf-2.index...  
Processing triggers for mime-support (3.59ubuntu1) ...  
Processing triggers for doc-base (0.10.7) ...  
Processing 1 added doc-base file...  
Setting up libpython3.5-minimal:amd64 (3.5.2-2ubuntu0~16.04.3) ...  
Setting up libpython3.5-stdlib:amd64 (3.5.2-2ubuntu0~16.04.3) ...  
Setting up libpython3.5:amd64 (3.5.2-2ubuntu0~16.04.3) ...  
Setting up python3.5-minimal (3.5.2-2ubuntu0~16.04.3) ...  
Setting up python3.5 (3.5.2-2ubuntu0~16.04.3) ...  
Setting up libexpat1-dev:amd64 (2.1.0-7ubuntu0.16.04.3) ...  
Setting up libpython3.5-dev:amd64 (3.5.2-2ubuntu0~16.04.3) ...  
Setting up libpython3-dev:amd64 (3.5.1-3) ...  
Setting up python3-pip-whl (8.1.1-2ubuntu0.4) ...  
Setting up python3.5-dev (3.5.2-2ubuntu0~16.04.3) ...  
Setting up python3-dev (3.5.1-3) ...  
Setting up python3-pip (8.1.1-2ubuntu0.4) ...  
Setting up python3-setuptools (20.7.0-1) ...  
Setting up python3-wheel (0.29.0-1) ...  
Processing triggers for libc-bin (2.23-0ubuntu9) ...  
wei@wei-VirtualBox:~$ pip3 install jupyter
```

(this installs jupyter. For more information on jupyter, please visit <http://jupyter.org/>. In a nutshell, it is a web application that allows you to create and share documents that contain live code, equations, visualizations and comments. It supports many programming languages including python).

```
wei@wei-VirtualBox: ~  
wei@wei-VirtualBox:~$ sudo apt-get install default-jre
```

(this installs java because we need scala)

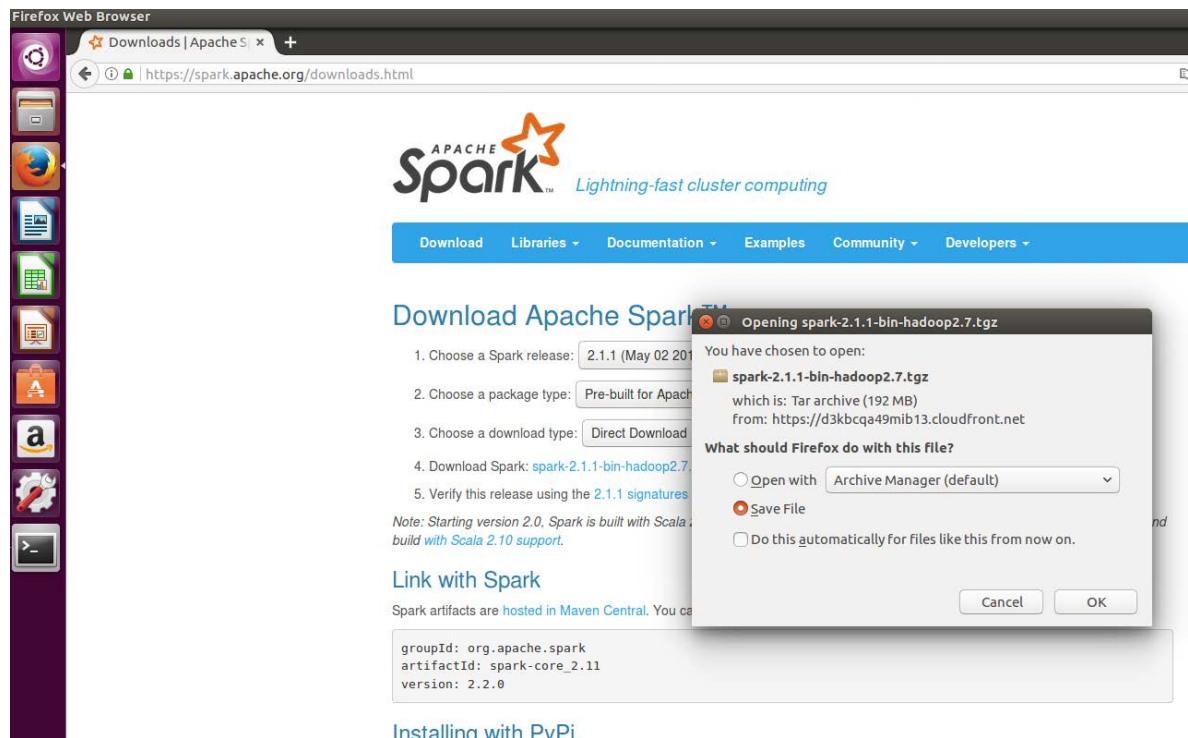
```
wei@wei-VirtualBox: ~  
provide /usr/bin/keytool (keytool) in auto mode  
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/jre/bin/jjs to provide /usr/bin/jjs (jjs) in auto mode  
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/jre/bin/pack200 to provide /usr/bin/pack200 (pack200) in auto mode  
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/jre/bin/rmiregistry to provide /usr/bin/rmiregistry (rmiregistry) in auto mode  
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/jre/bin/unpack200 to provide /usr/bin/unpack200 (unpack200) in auto mode  
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/jre/bin/orbd to provide /usr/bin/orbd (orbd) in auto mode  
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/jre/bin/servertool to provide /usr/bin/servertool (servertool) in auto mode  
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/jre/bin/tnameserv to provide /usr/bin/tnameserv (tnameserv) in auto mode  
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/jre/lib/jexec to provide /usr/bin/jexec (jexec) in auto mode  
Setting up default-jre-headless (2:1.8-56ubuntu2) ...  
Setting up openjdk-8-jre:amd64 (8u131-b11-2ubuntu1.16.04.3) ...  
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/jre/bin/policytool to provide /usr/bin/policytool (policytool) in auto mode  
Setting up default-jre (2:1.8-56ubuntu2) ...  
Processing triggers for libc-bin (2.23-0ubuntu9) ...  
wei@wei-VirtualBox:~$ sudo apt-get install scala
```

(this installs scala since we need to install Spark)

```
wei@wei-VirtualBox: ~  
Preparing to unpack .../scala-library_2.11.6-6_all.deb ...  
Unpacking scala-library (2.11.6-6) ...  
Selecting previously unselected package scala-parser-combinators.  
Preparing to unpack .../scala-parser-combinators_1.0.3-3_all.deb ...  
Unpacking scala-parser-combinators (1.0.3-3) ...  
Selecting previously unselected package scala-xml.  
Preparing to unpack .../scala-xml_1.0.3-3_all.deb ...  
Unpacking scala-xml (1.0.3-3) ...  
Selecting previously unselected package scala.  
Preparing to unpack .../scala_2.11.6-6_all.deb ...  
Unpacking scala (2.11.6-6) ...  
Processing triggers for doc-base (0.10.7) ...  
Processing 3 added doc-base files...  
Setting up libhwtjni-runtime-java (1.10-1) ...  
Setting up libjansi-native-java (1.0-4) ...  
Setting up libjansi-java (1.4-3) ...  
Setting up libjline2-java (2.11-4) ...  
Setting up scala-library (2.11.6-6) ...  
Setting up scala-parser-combinators (1.0.3-3) ...  
Setting up scala-xml (1.0.3-3) ...  
Setting up scala (2.11.6-6) ...  
update-alternatives: using /usr/share/scala-2.11/bin/scala to provide /usr/bin/scala (scala) in auto mode  
wei@wei-VirtualBox:~$ pip3 install py4j
```

(this installs py4j, Py4J enables Python programs running in a Python interpreter to dynamically access Java objects in a Java Virtual Machine).

Inside your virtual machine, use the browser to go to Spark website and download Spark 2.1.1.



Remember to move the downloaded file to home directory.

Then go back to your terminal and do following:

```
wei@wei-VirtualBox: ~  
wei@wei-VirtualBox:~$ pwd  
/home/wei  
wei@wei-VirtualBox:~$ sudo tar -zxvf spark-2.1.1-bin-hadoop2.7.tgz
```

(this unzips and installs the Spark and Hadoop)

After the unzip finishes, do following to set up the path correctly:

```
wei@wei-VirtualBox: ~  
wei@wei-VirtualBox:~$ export SPARK_HOME='home/ubuntu/spark-2.1.1-bin-hadoop2.7'  
wei@wei-VirtualBox:~$ export PATH=$SPARK_HOME:$PATH  
wei@wei-VirtualBox:~$ export PYTHONPATH=$SPARK_HOME/python:$PYTHONPATH  
wei@wei-VirtualBox:~$ export PYSPARK_DRIVER_PYTHON="jupyter"  
wei@wei-VirtualBox:~$ export PYSPARK_DRIVER_PYTHON_OPTS="notebook"  
wei@wei-VirtualBox:~$ export PYSPARK_PYTHON=python3
```

Now, we need to do something to avoid permission errorcd s:

```
wei@wei-VirtualBox: ~/spark-2.1.1-bin-hadoop2.7/python  
wei@wei-VirtualBox:~$ sudo chmod 777 spark-2.1.1-bin-hadoop2.7  
wei@wei-VirtualBox:~$ cd spark-2.1.1-bin-hadoop2.7/  
wei@wei-VirtualBox:~/spark-2.1.1-bin-hadoop2.7$ sudo chmod 777 python  
wei@wei-VirtualBox:~/spark-2.1.1-bin-hadoop2.7$ cd python/  
wei@wei-VirtualBox:~/spark-2.1.1-bin-hadoop2.7/python$ sudo chmod 777 pyspark  
wei@wei-VirtualBox:~/spark-2.1.1-bin-hadoop2.7/python$
```

In python folder, test your jupyter notebook.

```
wei@wei-VirtualBox: ~/spark-2.1.1-bin-hadoop2.7/python

wei@wei-VirtualBox:~$ sudo chmod 777 spark-2.1.1-bin-hadoop2.7
wei@wei-VirtualBox:~$ cd spark-2.1.1-bin-hadoop2.7/
wei@wei-VirtualBox:~/spark-2.1.1-bin-hadoop2.7$ sudo chmod 777 python
wei@wei-VirtualBox:~/spark-2.1.1-bin-hadoop2.7$ cd python/
wei@wei-VirtualBox:~/spark-2.1.1-bin-hadoop2.7/python$ sudo chmod 777 pyspark
wei@wei-VirtualBox:~/spark-2.1.1-bin-hadoop2.7/python$ jupyter notebook
```

And you should see jupyter notebook shows up in your browser.

3. Some extra steps that allow you to import pyspark from any directory.

```
wei@wei-VirtualBox: ~/spark-2.1.1-bin-hadoop2.7

wei@wei-VirtualBox:~$ pip3 install findspark
Collecting findspark
  Downloading findspark-1.1.0-py2.py3-none-any.whl
Installing collected packages: findspark
Successfully installed findspark-1.1.0
You are using pip version 8.1.1, however version 9.0.1 is available.
You should consider upgrading via the 'pip install --upgrade pip' command.
wei@wei-VirtualBox:~$ ls
Desktop      examples.desktop  Public          Templates
Documents    Music             spark-2.1.1-bin-hadoop2.7  Videos
Downloads    Pictures          spark-2.1.1-bin-hadoop2.7.tgz
wei@wei-VirtualBox:~$ cd s
bash: cd: s: No such file or directory
wei@wei-VirtualBox:~$ cd spark-2.1.1-bin-hadoop2.7/
wei@wei-VirtualBox:~/spark-2.1.1-bin-hadoop2.7$ python3
```

```
wei@wei-VirtualBox: ~/spark-2.1.1-bin-hadoop2.7
wei@wei-VirtualBox:~/spark-2.1.1-bin-hadoop2.7$ pwd
/home/wei/spark-2.1.1-bin-hadoop2.7
wei@wei-VirtualBox:~/spark-2.1.1-bin-hadoop2.7$ python3
Python 3.5.2 (default, Sep 14 2017, 22:51:06)
[GCC 5.4.0 20160609] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> import findspark

>>> findspark.init('/home/wei/spark-2.1.1-bin-hadoop2.7')
>>> 
```

Now, you should have set up your system on virtual machine. From your terminal start jupyter notebook and start working with Python.