

Apache Spark with Vagrant

1. Download and Install [Vagrant](#)
2. Download and Install [VirtualBox](#)
3. Create a new directory and name it `pyspark`
4. Download `vagrantfile` from our [googledrive](#) and save it in `pyspark` directory
5. Open up your shell terminal and change the working directory to `pyspark`

6. Type in,

```
$ vagrant up
```

This command will initiate installation of Anaconda, Java EMR, and Spark. This process will take up to 20 minutes.

7. Once installation finishes, type in

```
$ vagrant ssh
```

This command will log you into your virtual VM

8. In the root directory, type in

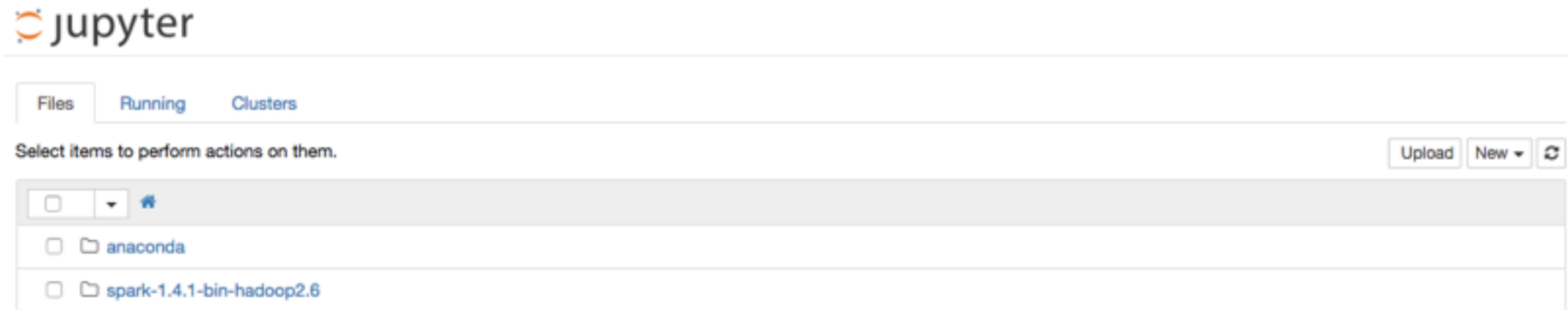
```
vagrant@precise64:~$ IPYTHON_OPTS="notebook --ip=0.0.0.0  
--no-browser" ./spark-1.4.1-bin-hadoop2.6/bin/pyspark
```

This launches ipython notebook with
Apache pyspark API

9. Open up your web browser and type in

<http://localhost:8888/tree>

10. You should see the following in your browser



11. Click on New button in the top right corner to start a new notebook

12. In the new notebook, run the following python script to verify correct installation of spark

```
In [3]: print "Spark Version", sc.version  
Spark Version 1.4.1
```

```
In [ ]:
```

13. You can also upload your own `.ipynb` file by clicking on Upload



14. When finished, exit out of the VM shell by typing

```
vagrant@precise64:~$ exit
```

and in your local console, type

```
$ vagrant halt
```

this will shut down the VM.