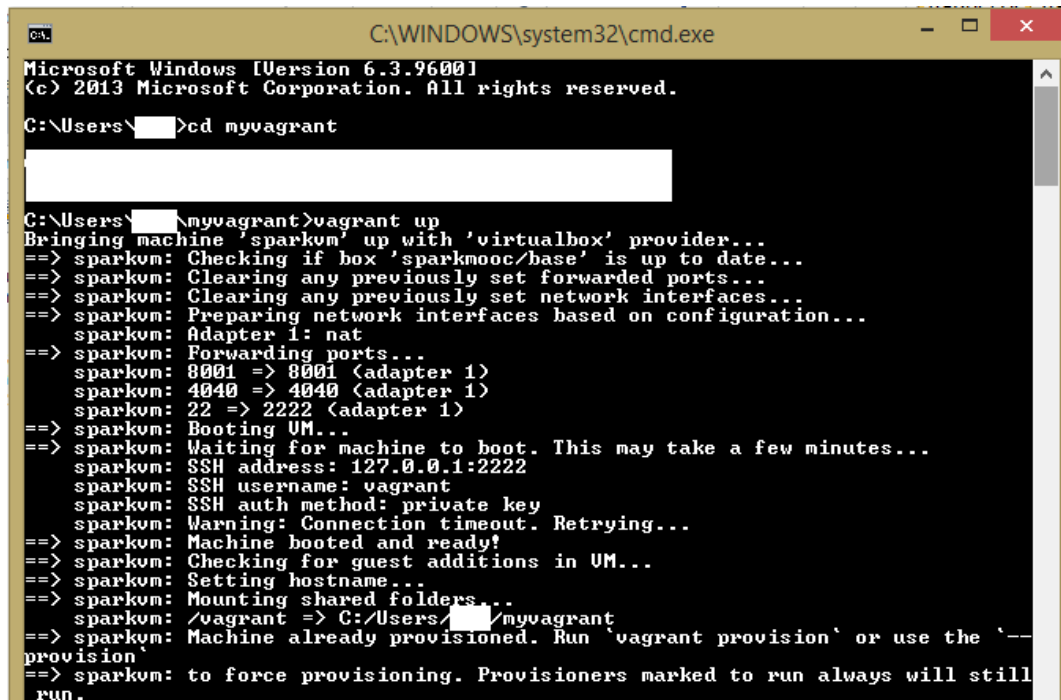


How to upload data files to Vagrant

Tags: Spark, VirtualBox, Vagrant, Jupyter

Usually the upload feature in the Jupyter window fails when data files are large. I will demonstrate a method that uses shared folder. You should know the shared folder's name in your hard drive. You can get it from the output when you start vagrant:

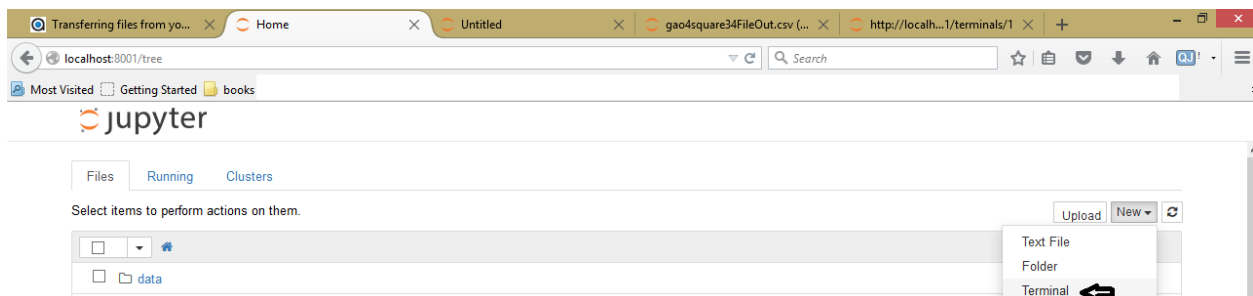


```
C:\WINDOWS\system32\cmd.exe
Microsoft Windows [Version 6.3.9600]
(c) 2013 Microsoft Corporation. All rights reserved.

C:\Users\>cd myvagrant

C:\Users\>vagrant up
Bringing machine 'sparkvm' up with 'virtualbox' provider...
==> sparkvm: Checking if box 'sparkmooc/base' is up to date...
==> sparkvm: Clearing any previously set forwarded ports...
==> sparkvm: Clearing any previously set network interfaces...
==> sparkvm: Preparing network interfaces based on configuration...
    sparkvm: Adapter 1: nat
==> sparkvm: Forwarding ports...
    sparkvm: 8001 => 8001 (adapter 1)
    sparkvm: 4040 => 4040 (adapter 1)
    sparkvm: 22 => 2222 (adapter 1)
==> sparkvm: Booting VM...
==> sparkvm: Waiting for machine to boot. This may take a few minutes...
    sparkvm: SSH address: 127.0.0.1:2222
    sparkvm: SSH username: vagrant
    sparkvm: SSH auth method: private key
    sparkvm: Warning: Connection timeout. Retrying...
==> sparkvm: Machine booted and ready!
==> sparkvm: Checking for guest additions in VM...
==> sparkvm: Setting hostname...
==> sparkvm: Mounting shared folders...
    sparkvm: /vagrant => C:/Users/ /myvagrant
==> sparkvm: Machine already provisioned. Run 'vagrant provision' or use the '--
provision'
==> sparkvm: to force provisioning. Provisioners marked to run always will still
run.
```

Typically, the shared folder is c:/Users/USERNAME/myvagrant. Copy the file containing your dataset in this directory. The shared folder is merely a folder in your virtual box that is running a linux OS. You can explore the shared folder using a terminal in the Jupyter.





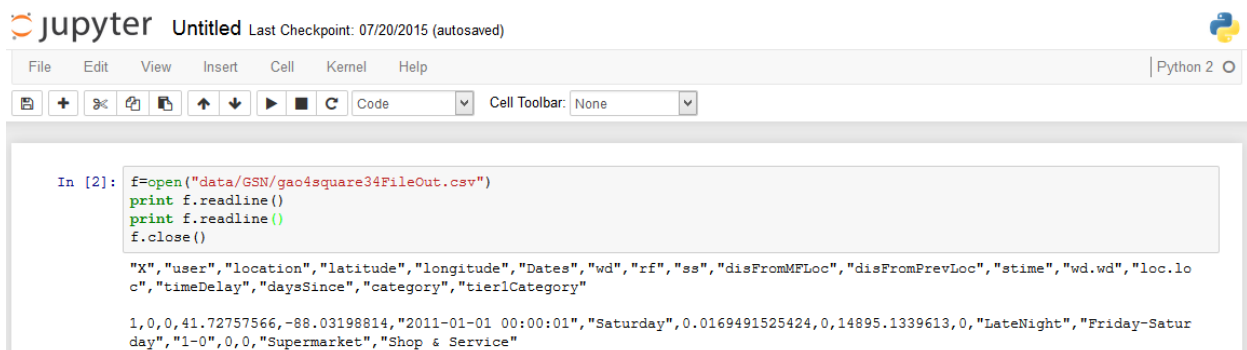
```
$ ls
bin  dev  home  lib      media  opt  root  sbin  sys  usr  var  wheelhouse
boot etc  initrd.img  lost+found  mnt  proc  run  srv  tmp  vagrant  vmlinuz
$ ls vagrant
Vagrantfile  gao4square34FileOut.csv  myvagrantroot.txt
$
```

As you can see, I have uploaded the file `gao4square34FileOut.csv` which over 300MB. Now I have to copy this file into `/home/vagrant` folder so that I can see the dataset from Jupyter.



```
$ cp vagrant/gao4square34FileOut.csv home/vagrant/data/GSN
$ cd home/vagrant/data/GSN
$ head -n 3 gao4square34FileOut.csv
"x","user","location","latitude","longitude","Dates","wd","rf","ss","
"daysSince","category","tier1Category"
1,0,0,41.72757566,-88.03198814,"2011-01-01 00:00:01","Saturday",0.016
"Supermarket","Shop & Service"
2,0,1245,41.94358766,-87.64941126,"2011-01-01 03:47:42","Saturday",0.
"Saturday-Saturday","0-1245",13661,0,"Gay Bar","Nightlife Spot"
$
```

Some of the output is cut due to the space constraint. Now let's write a python program to read the first two lines of the file.



The dataset is now in `vagrant` and we can read it.

Another way of uploading dataset is to use `scp` command. Format of the command is following:

```
scp fileYouWantToUpload vagrant@127.0.0.1:2222
```

You are likely to get an error saying bad file number. This is due to the fact that SSH is typically blocked on the specified port. You can resolve the issue by editing SSH configuration file. You can follow this [post](#).

Thanks for reading!