CS109/Stat121/AC209/E-109
Data Science

Using Apache Spark with Vagrant

Vagrant and Virtual Box

Vagrant is a wrapper around virtualization software such as VirtualBox.

Due to the complexity involved in installing and configuring Apache Spark, we will use Vagrant and VirtualBox to install Spark and Anaconda distribution inside a virtual machine environment running on Ubuntu Linux.

Vagrant and Virtual Box

Our Vagrant wrapper will use 2 processors and approximately 1.0 GB of memory. It's important that you close down other work to release these resources for Vagrant.

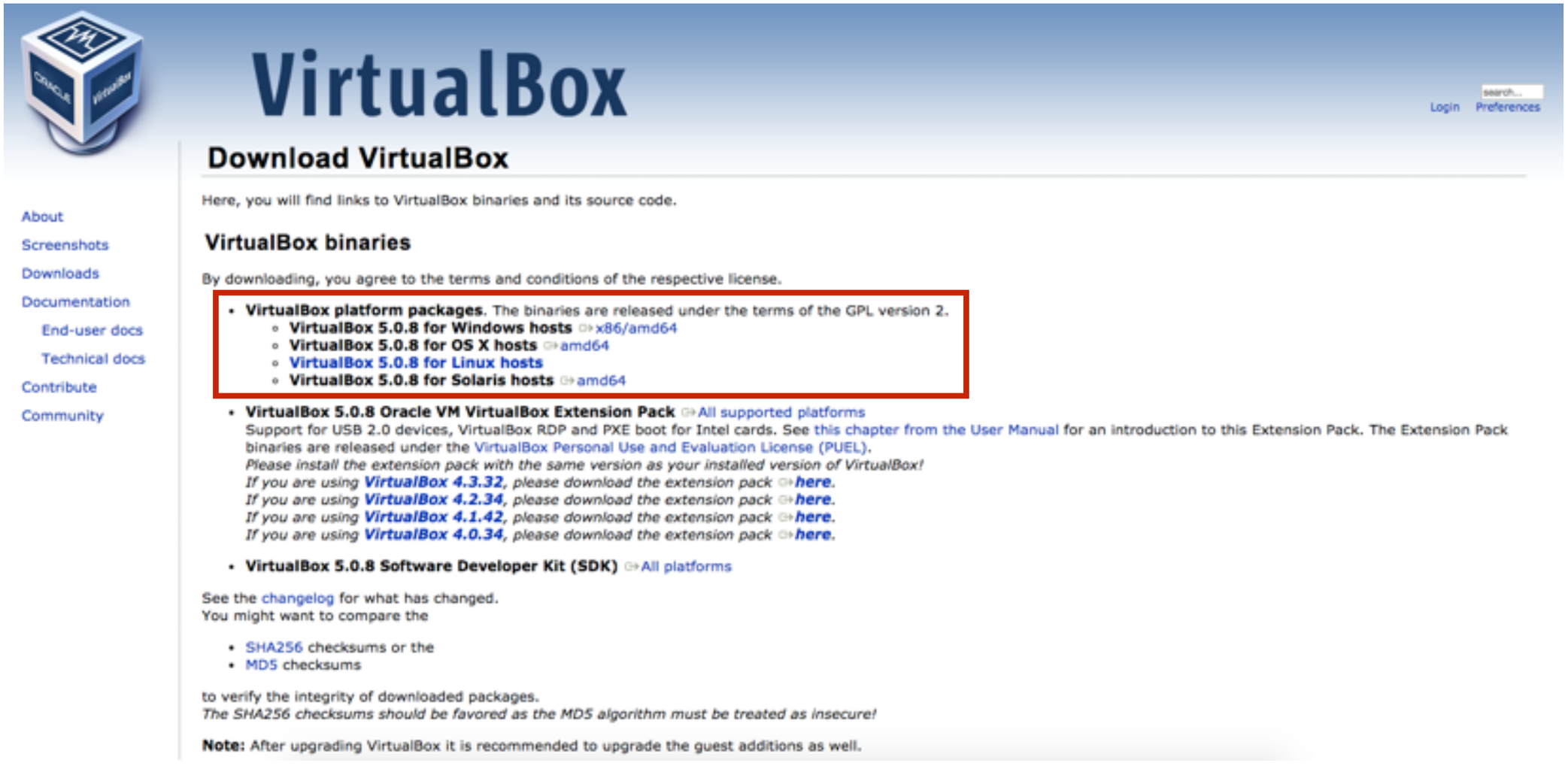
We will be using port 8888 to interact with the virtual machine; therefore, it is a good idea to save and close existing ipython notebooks before we begin.

1. Download and Install Vagrant

https://www.vagrantup.com/downloads.html

2. Download and Install Virtual Box

https://www.virtualbox.org/wiki/Downloads



VirtualBox

Login Preferences

Download VirtualBox

About
Screenshots
Downloads
Documentation
End-user docs
Technical docs
Contribute
Community

Here, you will find links to VirtualBox binaries and its source code.

VirtualBox binaries

By downloading, you agree to the terms and conditions of the respective license.

- **VirtualBox platform packages**. The binaries are released under the terms of the GPL version 2.
  - **VirtualBox 5.0.8 for Windows hosts** x86/amd64
  - **VirtualBox 5.0.8 for OS X hosts** amd64
  - **VirtualBox 5.0.8 for Linux hosts**
  - **VirtualBox 5.0.8 for Solaris hosts** amd64
- **VirtualBox 5.0.8 Oracle VM VirtualBox Extension Pack** All supported platforms
  Support for USB 2.0 devices, VirtualBox RDP and PXE boot for Intel cards. See this chapter from the User Manual for an introduction to this Extension Pack. The Extension Pack binaries are released under the VirtualBox Personal Use and Evaluation License (PUEL).
  Please install the extension pack with the same version as your installed version of VirtualBox!
  *If you are using **VirtualBox 4.3.32**, please download the extension pack **here**.*
  *If you are using **VirtualBox 4.2.34**, please download the extension pack **here**.*
  *If you are using **VirtualBox 4.1.42**, please download the extension pack **here**.*
  *If you are using **VirtualBox 4.0.34**, please download the extension pack **here**.*
- **VirtualBox 5.0.8 Software Developer Kit (SDK)** All platforms

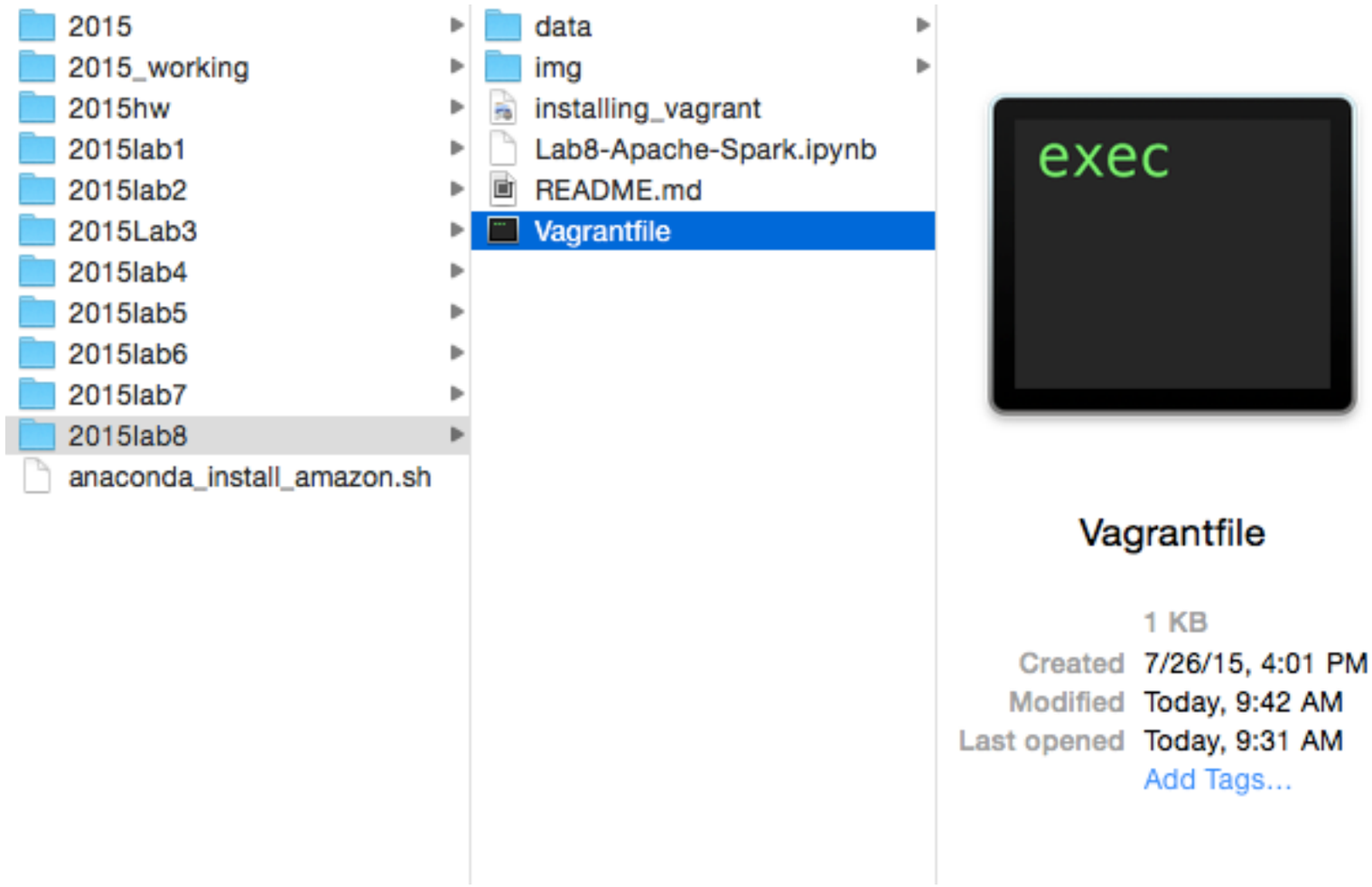
See the changelog for what has changed.
You might want to compare the

- SHA256 checksums or the
- MD5 checksums

to verify the integrity of downloaded packages.
The SHA256 checksums should be favored as the MD5 algorithm must be treated as insecure!

Note: After upgrading VirtualBox it is recommended to upgrade the guest additions as well.

3. Inside your lab folder your should see a file called `Vagrantfile`



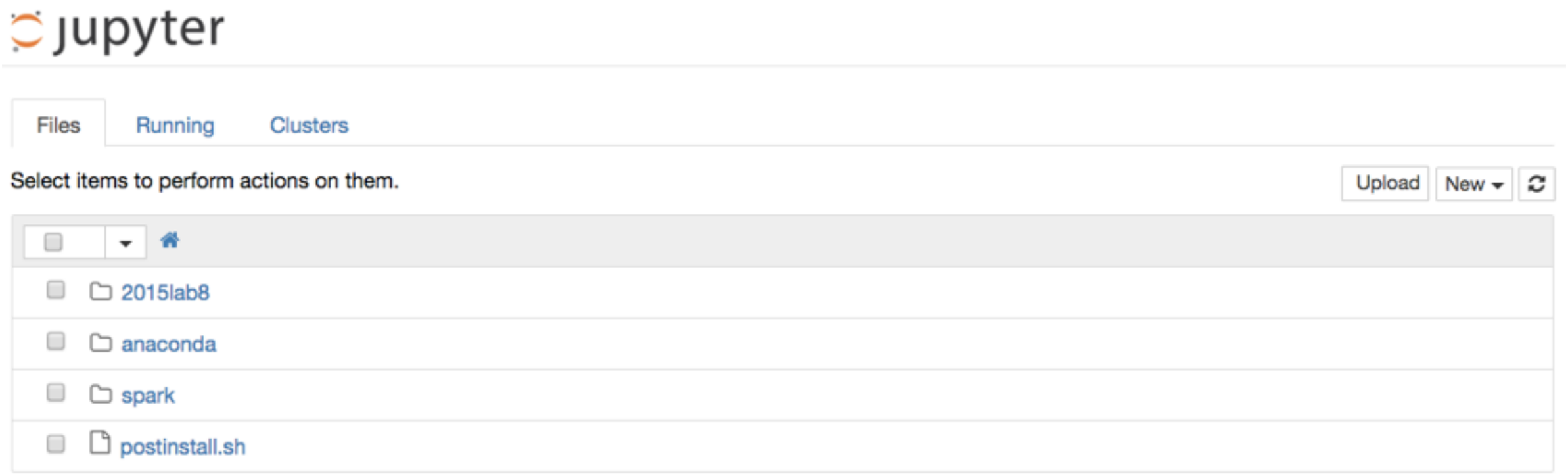
4. Inside the `2015lab8/` directory, type in

```
$ vagrant up
```

When you run this command for the first time, Vagrant will install Anaconda, Java EMR, and Apache Spark. This process may take up to 25 minutes. In subsequent runs, it should only take less than a minute.

5. Open up your web browser and type in the url:

`http://localhost:8888/tree`



6. Open up `2015lab8/` folder

`2015lab8/` folder has been synced to the `2015lab8/` folder in your local machine. Any changes that you make here will be reflected in the same folder in your local machine.

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

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

and in your local machine command line, inside `2015lab8/`, type

```
$ vagrant halt
```

This will shut down the VM.