

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>



The screenshot shows the Vagrant website's download page. The header includes the Vagrant logo and navigation links: VMWARE INTEGRATION, DOWNLOADS, DOCUMENTATION, BLOG, and ABOUT. A left sidebar with a blue background contains the word 'DOWNLOAD' and two links: 'Latest' and 'Old Versions'. The main content area is titled 'DOWNLOAD VAGRANT' and contains a paragraph explaining that the following are downloads for the latest version (1.7.4), with links to SHA256 checksums and the changelog. Below this, three operating system options are listed, each with its logo and download links: MAC OS X (Universal 32 and 64-bit), WINDOWS (Universal 32 and 64-bit), and LINUX (DEB) (32-bit and 64-bit).

VAGRANT

VMWARE INTEGRATION DOWNLOADS DOCUMENTATION BLOG ABOUT

DOWNLOAD

[Latest](#)

[Old Versions](#)

DOWNLOAD VAGRANT

Below are all available downloads for the latest version of Vagrant (1.7.4). Please download the proper package for your operating system and architecture. You can find SHA256 checksums for packages [here](#), and you can find the version changelog [here](#).

 **MAC OS X**
[Universal \(32 and 64-bit\)](#)

 **WINDOWS**
[Universal \(32 and 64-bit\)](#)

 **LINUX (DEB)**
[32-bit](#) | [64-bit](#)

2. Download and Install Virtual Box

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



The screenshot shows the 'Download VirtualBox' page on the official VirtualBox website. The page has a blue header with the VirtualBox logo and navigation links. The main content area is white with a red border around the 'VirtualBox binaries' section. The 'VirtualBox binaries' section lists the following:

- **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](#)

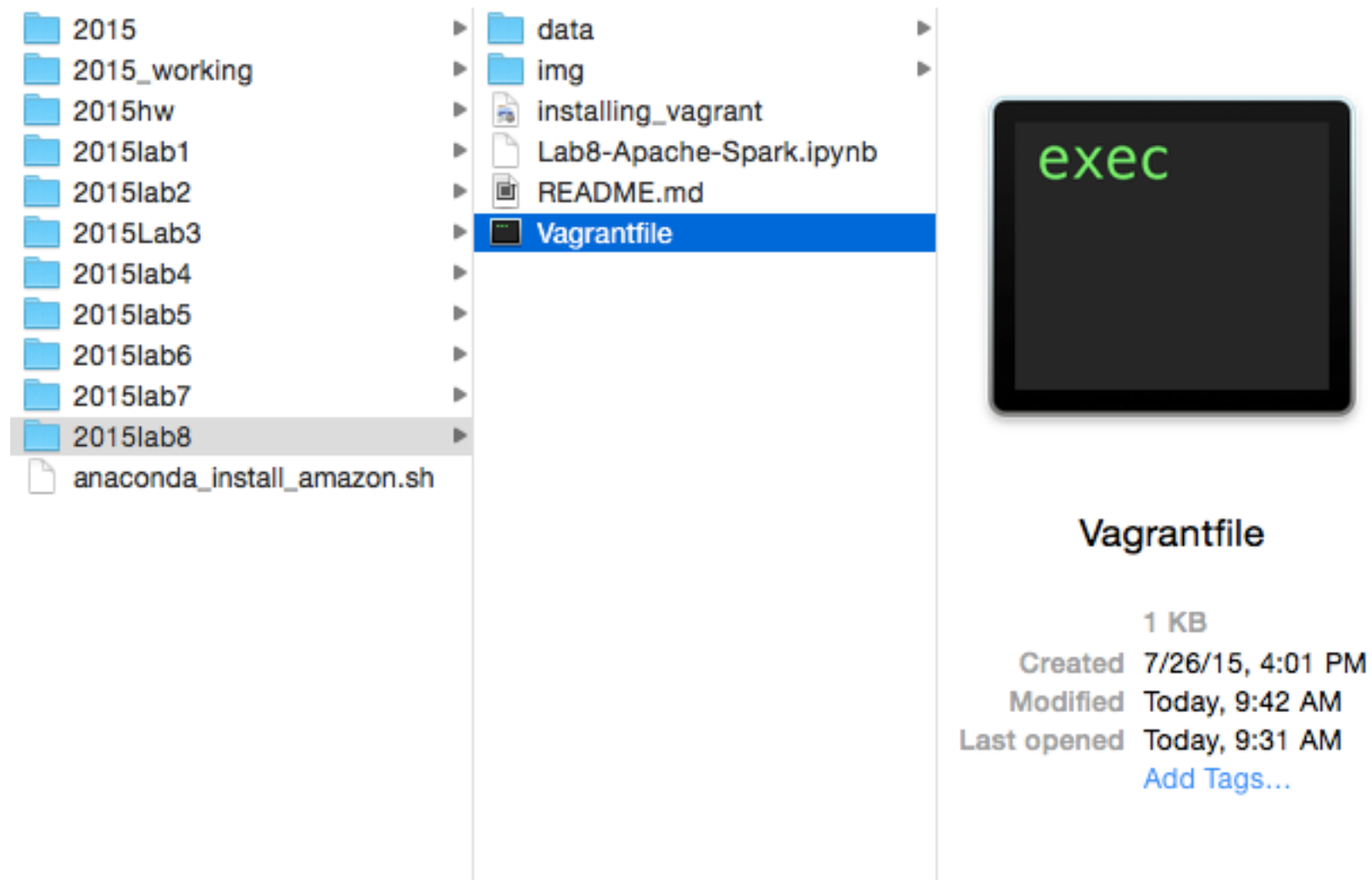
See the [changelog](#) for what has changed.
You might want to compare the

- [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 you should see a file called **Vagrantfile**



4. Inside the `20151ab8/` 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>



Files

Running

Clusters

Select items to perform actions on them.

Upload

New ▾



<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
<input type="checkbox"/>		2015lab8	
<input type="checkbox"/>		anaconda	
<input type="checkbox"/>		spark	
<input type="checkbox"/>		postinstall.sh	

6. Open up 2015lab8/ folder



Files Running Clusters

Select items to perform actions on them. Upload New ▾ ↺

☐ ▾

🏠 / 2015lab8

<input type="checkbox"/>	..
<input type="checkbox"/>	data
<input type="checkbox"/>	img
<input type="checkbox"/>	pyspark
<input checked="" type="checkbox"/>	Lab8-Apache-Spark-Vagrant.ipynb
<input type="checkbox"/>	README.md
<input type="checkbox"/>	installing_vagrant.pdf

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.