

# Apache Spark Lab


- You will need to obtain a directory of files to run this lab. The files are on my github account.
- It is assumed you have Virtual Box installed
- Install Vagrant
- You will need to know how to open a terminal window on your particular pc

## Install Vagrant

Follow the instruction here for you particular machine




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


<https://github.com/marilynwaldman/cuSpark>







This repository   Search

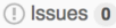
Pull requests   Issues   Gist


  


 marilynwaldman / cuSpark


 1    0    0

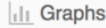
 <> Code


 Issues 0

 Pull requests 0


 Wiki


 Pulse


 Graphs


 Settings

No description or website provided. — Edit

 7 commits

 1 branch

 0 releases

 1 contributor

Branch: master ▾

New pull request


New file


Upload files

Find file


SSH ▾

git@github.com:marilynwal
















Download ZIP


 marilynwaldman fixed gitignore


Latest commit 5d9dd3e 4 days ago


 <a href="#">.gitignore</a>	fixed gitignore	4 days ago
 <a href="#">README.md</a>	updated readme	4 days ago
 <a href="#">Vagrantfile</a>	first push	15 days ago
 <a href="#">beowulf.txt</a>	total rewrite	4 days ago
 <a href="#">hamlet.txt</a>	total rewrite	4 days ago
 <a href="#">lab1-map-reduce.ipynb</a>	total rewrite	4 days ago
 <a href="#">lab2_sparkRDDs.ipynb</a>	total rewrite	4 days ago
 <a href="#">lab3-sparkWordCount.ipynb</a>	total rewrite	4 days ago
 <a href="#">lab4-moreSpark.ipynb</a>	total rewrite	4 days ago
 <a href="#">labfile</a>	total rewrite	4 days ago

You can download the repo with Git or  
download the directory with  
"Download ZIP"

 7 commits

 1 branch

 0 releases

 1 contributor

Branch: **master**  [New pull request](#)[New file](#) [Upload files](#) [Find file](#) [SSH](#)  `git@github.com:marilynwal`   [Download ZIP](#) **marilynwaldman** fixed gitignore Latest commit 5d9dd3e 4 days ago [.aitianore](#) fixed aitianore 4 days ago











## With Git

```
cuSpark — -bash — 170x32
Last login: Sun Apr 17 11:12:48 on ttys001
[Marilyns-MacBook-Pro:~ marilynwaldman$ git clone github.com:marilynwaldman/cuSpark.git
Cloning into 'cuSpark'...
Warning: Permanently added the RSA host key for IP address '192.30.252.121' to the list of known hosts.
remote: Counting objects: 56, done.
remote: Compressing objects: 100% (16/16), done.
remote: Total 56 (delta 3), reused 0 (delta 0), pack-reused 32
Receiving objects: 100% (56/56), 1.86 MiB | 739.00 KiB/s, done.
Resolving deltas: 100% (9/9), done.
Checking connectivity... done.
[Marilyns-MacBook-Pro:~ marilynwaldman$ cd cuSpark
Marilyns-MacBook-Pro:cuSpark marilynwaldman$
```

**notice directory name is cuSpark**

OR

DOWNLOAD and unzip

View Action Arrange By Share Edit Tags					Search	
cuSpark-master					+	
Name	^	Date Modified	Size	Kind		
 .gitignore		Apr 15, 2016, 10:54 AM	20 bytes	Unix e...cutable		
 beowulf.txt		Apr 15, 2016, 10:54 AM	300 KB	Plain Text		
 hamlet.txt		Apr 15, 2016, 10:54 AM	153 KB	Plain Text		
 lab1-map-reduce.ipynb		Apr 15, 2016, 10:54 AM	13 KB	TextM...cument		
 lab2_sparkRDDs.ipynb		Apr 15, 2016, 10:54 AM	15 KB	TextM...cument		
 lab3-sparkWordCount.ipynb		Apr 15, 2016, 10:54 AM	64 KB	TextM...cument		
 lab4-moreSpark.ipynb		Apr 15, 2016, 10:54 AM	8 KB	TextM...cument		
 logfile		Apr 15, 2016, 10:54 AM	32 KB	Unix e...cutable		
 README.md		Apr 15, 2016, 10:54 AM	722 bytes	Markd...cument		
 Vagrantfile		Apr 15, 2016, 10:54 AM	726 bytes	Unix e...cutable		

Open a terminal window and change  
directory to **cuSpark-master**

▼	cuSpark	Today, 9:21 AM	--	Folder	Today, 9:21 AM
	beowulf.txt	Today, 9:21 AM	300 KB	Plain Text	Today, 9:21 AM
	hamlet.txt	Today, 9:21 AM	153 KB	Plain Text	Today, 9:21 AM
	lab1-map-reduce.ipynb	Today, 9:21 AM	13 KB	TextM...cument	Today, 9:21 AM
	lab2_sparkRDDs.ipynb	Today, 9:21 AM	15 KB	TextM...cument	Today, 9:21 AM
	lab3-sparkWordCount.ipynb	Today, 9:21 AM	64 KB	TextM...cument	Today, 9:21 AM
	lab4-moreSpark.ipynb	Today, 9:21 AM	8 KB	TextM...cument	Today, 9:21 AM
	logfile	Today, 9:21 AM	32 KB	TextEd...ument	Today, 9:21 AM
	README.md	Today, 9:21 AM	722 bytes	Markd...cument	Today, 9:21 AM
	Vagrantfile	Today, 9:21 AM	726 bytes	Unix e...cutable	Today, 9:21 AM

We will upload all files except README and Vagrantfile



Open a terminal window and change directory to  
either **cuSpark** or **cuSpark-master** depending  
on how you obtained your files

```
Last login: Tue Apr 19 09:20:44 on ttys000
Marilyns-MacBook-Pro:~ marilynwaldman$ cd cuSpark
Marilyns-MacBook-Pro:cuSpark marilynwaldman$ ls -l
total 1184
-rw-r--r--  1 marilynwaldman  staff    722 Apr 19 09:21 README.md
-rwxr-xr-x  1 marilynwaldman  staff    726 Apr 19 09:21 Vagrantfile
-rw-r--r--  1 marilynwaldman  staff 299992 Apr 19 09:21 beowulf.txt
-rw-r--r--  1 marilynwaldman  staff 152793 Apr 19 09:21 hamlet.txt
-rw-r--r--  1 marilynwaldman  staff  12608 Apr 19 09:21 lab1-map-reduce.ipynb
-rw-r--r--  1 marilynwaldman  staff  14752 Apr 19 09:21 lab2_sparkRDDs.ipynb
-rw-r--r--  1 marilynwaldman  staff  64289 Apr 19 09:21 lab3-sparkWordCount.ipynb
-rw-r--r--  1 marilynwaldman  staff   7514 Apr 19 09:21 lab4-moreSpark.ipynb
-rw-r--r--  1 marilynwaldman  staff  32085 Apr 19 09:21 logfile
Marilyns-MacBook-Pro:cuSpark marilynwaldman$
```

## From the terminal window issue "**vagrant up**"

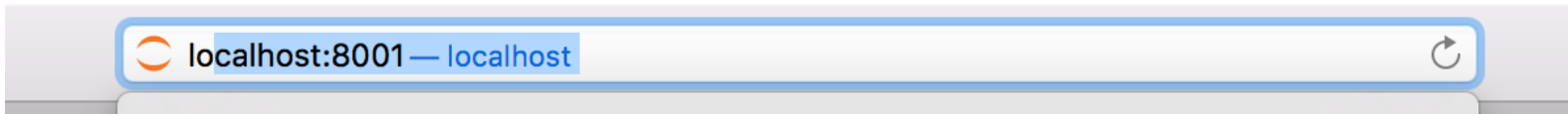
```
Marilyns-MacBook-Pro:cuSpark marilynwaldman$ vagrant up
Bringing machine 'cuSparkVM' up with 'virtualbox' provider...
==> cuSparkVM: Importing base box 'sparkmooc/base'...
==> cuSparkVM: Matching MAC address for NAT networking...
==> cuSparkVM: Checking if box 'sparkmooc/base' is up to date...
==> cuSparkVM: Setting the name of the VM: cuSparkVM
==> cuSparkVM: Clearing any previously set network interfaces...
==> cuSparkVM: Preparing network interfaces based on configuration...
    cuSparkVM: Adapter 1: nat
==> cuSparkVM: Forwarding ports...
    cuSparkVM: 8001 (guest) => 8001 (host) (adapter 1)
    cuSparkVM: 4040 (guest) => 4040 (host) (adapter 1)
    cuSparkVM: 22 (guest) => 2222 (host) (adapter 1)
==> cuSparkVM: Booting VM...
==> cuSparkVM: Waiting for machine to boot. This may take a few minutes...
    cuSparkVM: SSH address: 127.0.0.1:2222
    cuSparkVM: SSH username: vagrant
    cuSparkVM: SSH auth method: private key
    cuSparkVM:
    cuSparkVM: Vagrant insecure key detected. Vagrant will automatically replace
    cuSparkVM: this with a newly generated keypair for better security.
    cuSparkVM:
    cuSparkVM: Inserting generated public key within guest...
    cuSparkVM: Removing insecure key from the guest if it's present...
    cuSparkVM: Key inserted! Disconnecting and reconnecting using new SSH key...
==> cuSparkVM: Machine booted and ready!
==> cuSparkVM: Checking for guest additions in VM...
    cuSparkVM: The guest additions on this VM do not match the installed version of
    cuSparkVM: VirtualBox! In most cases this is fine, but in rare cases it can
    cuSparkVM: prevent things such as shared folders from working properly. If you see
    cuSparkVM: shared folder errors, please make sure the guest additions within the
    cuSparkVM: virtual machine match the version of VirtualBox you have installed on
    cuSparkVM: your host and reload your VM.
    cuSparkVM:
    cuSparkVM: Guest Additions Version: 4.3.10
    cuSparkVM: VirtualBox Version: 5.0
==> cuSparkVM: Setting hostname...
==> cuSparkVM: Mounting shared folders...
    cuSparkVM: /vagrant => /Users/marilynwaldman/cuSpark
Marilyns-MacBook-Pro:cuSpark marilynwaldman$
```

***Point your browser to***

`http://127.0.0.1:8001`

**or**

`localhost:8001`





Files

Running

Clusters

Select items to perform actions on them.

Upload

New ▾



☐  data

☐  spark\_mooc\_version

☐  spark\_notebook.py

# Upload all files except Vagrantfile and README.md



Files

Running

Clusters

Select items to perform actions on them.

Upload

New ▾

☐

▼

☐

data

☐

spark\_mooc\_version

☐

spark\_notebook.py

Click to browse for a file to up

Files

Running

Clusters

Select items to perform actions on them.

Upload

New ▾

☐

▼

☐

data

☐

spark\_mooc\_version

☐

spark\_notebook.py

Click to browse for a file to up



# Upload all files

Select items to perform actions on them.

Upload New ↕

<input type="checkbox"/>	▼	🏠		
<input type="checkbox"/>		logfile	Upload	Cancel
<input type="checkbox"/>		lab4-moreSpark.ipynb	Upload	Cancel
<input type="checkbox"/>		lab3-sparkWordCount.ipynb	Upload	Cancel
<input type="checkbox"/>		lab2_sparkRDDs.ipynb	Upload	Cancel
<input type="checkbox"/>		lab1-map-reduce.ipynb	Upload	Cancel
<input type="checkbox"/>		hamlet.txt	Upload	Cancel
<input type="checkbox"/>		beowulf.txt	Upload	Cancel
<input type="checkbox"/>		data		
<input type="checkbox"/>		spark_mooc_version		
<input type="checkbox"/>		spark_notebook.py		

# Choose lab1



Files

Running

Clusters

Select items to perform actions on them.

Upload

New ▾



<input type="checkbox"/>	<input type="button" value="v"/>	
<input type="checkbox"/>		data
<input type="checkbox"/>		lab1-map-reduce.ipynb
<input type="checkbox"/>		lab2_sparkRDDs.ipynb
<input type="checkbox"/>		lab3-sparkWordCount.ipynb
<input type="checkbox"/>		lab4-moreSpark.ipynb
<input type="checkbox"/>		beowulf.txt
<input type="checkbox"/>		hamlet.txt
<input type="checkbox"/>		logfile
<input type="checkbox"/>		spark_mooc_version
<input type="checkbox"/>		spark_notebook.py





# Introduction to the Map and Reduce Abstractions

## The `map` abstraction

```
In [1]: #Initialize data
nums = range(10)
print nums
print type(nums)

[0, 1, 2, 3, 4, 5, 6, 7, 8, 9]
<type 'list'>
```

```
In [20]: #Python only

map(lambda x: x*x, nums)
```

```
Out[20]: [0, 1, 4, 9, 16, 25, 36, 49, 64, 81]
```

```
In [27]: #Spark - push nums list onto five executors
sparkNums = sc.parallelize(nums, 5)
#map - this is a transformation
squares = sparkNums.map(lambda x: x*x)
#print result - this is an action - push results back to the driver
print squares.collect()
```

Select items to perform actions on them.

Upload New ↻

<input type="checkbox"/>	<input type="text"/>	
<input type="checkbox"/>	data	
<input type="checkbox"/>	lab1-map-reduce.ipynb	Running
<input type="checkbox"/>	lab2_sparkRDDs.ipynb	
<input type="checkbox"/>	lab3-sparkWordCount.ipynb	
<input type="checkbox"/>	lab4-moreSpark.ipynb	
<input type="checkbox"/>	beowulf.txt	
<input type="checkbox"/>	hamlet.txt	
<input type="checkbox"/>	logfile	
<input type="checkbox"/>	spark_mooc_version	
<input type="checkbox"/>	spark_notebook.py	

Only one notebook may be running at a time. Shutdown lab 1 before starting lab 2

Files

Running

Clusters

Duplicate

Shutdown

Upload

New

1

<input type="checkbox"/>	data	
<input checked="" type="checkbox"/>	lab1-map-reduce.ipynb	Running
<input type="checkbox"/>	lab2_sparkRDDs.ipynb	
<input type="checkbox"/>	lab3-sparkWordCount.ipynb	
<input type="checkbox"/>	lab4-moreSpark.ipynb	
<input type="checkbox"/>	beowulf.txt	