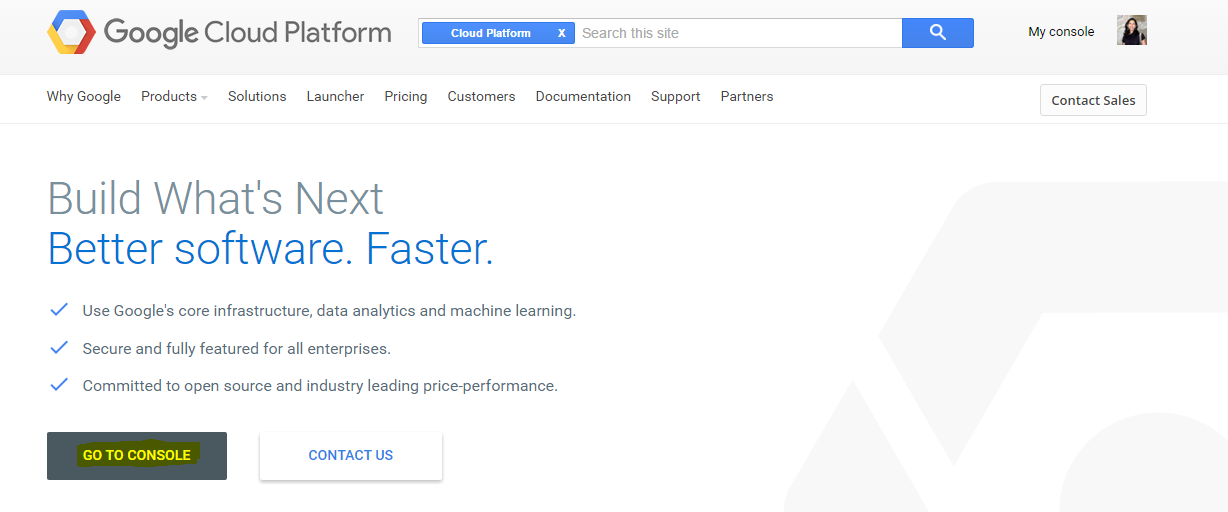
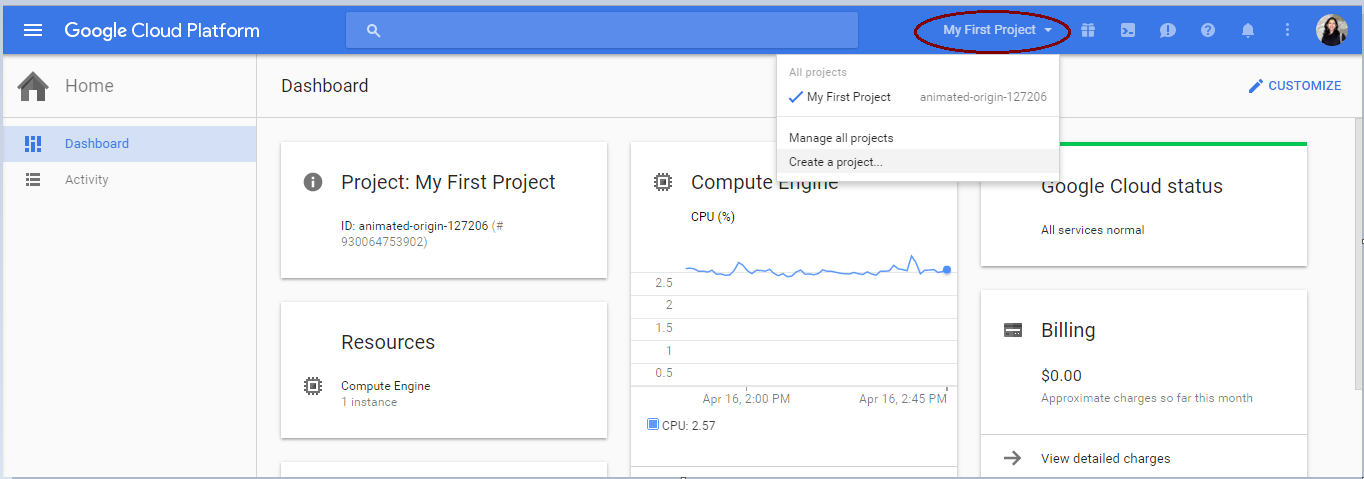
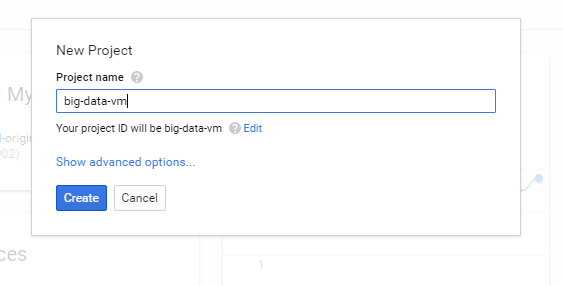
# **ALTERNATIVE SOLUTIONS FOR INSTALLING AND RUNNING SPARK**

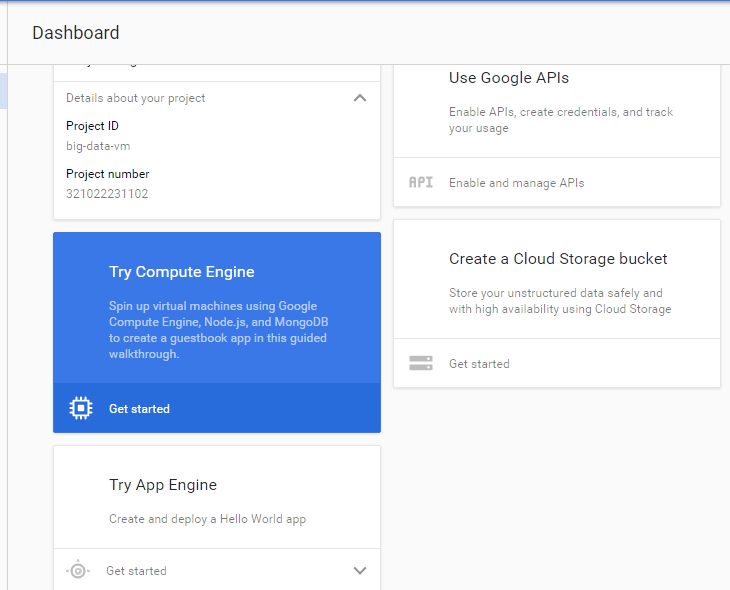
***NOTE***: With Amazon’s solution, you will have one year’s worth of free cloud services, but limited options with the size and memory of your virtual machine. You need to be careful that all the configurations you choose belong to the free tier (will be highlighted). With Google, you have $300 worth of services free for a period of 2 months beyond which charges, depending on your configurations, will be levied.

## ALTERNATIVE 1: Installing Ubuntu and Spark on Google Cloud’s Virtual Machine

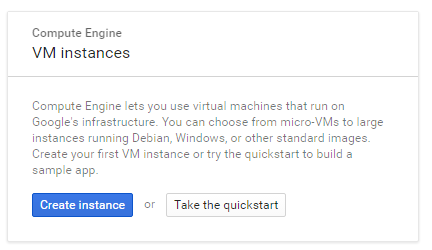
1. Go to <https://cloud.google.com/>
2. Sign-up with your Gmail id. You will have to provide your credit card information among others. Currently, Google provides $300 worth of cloud services free for two months. You may be charged as per the services you have deployed after the 2-month period. Please check their pricing scheme for more. For our assignment purposes, we would need the bare minimum and this may amount to ~ $1 per day beyond the 2-month period.
3. Once you have registered, click on “Go to console”
4. On the screen, click on My First Project on the top right of the window and select Create a Project.



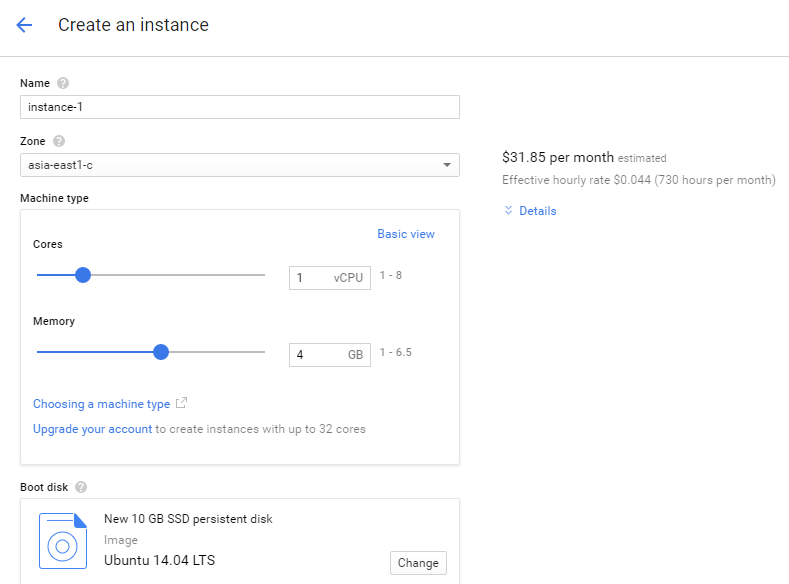
1. Give a suitable name for your project (example: big-data-vm) and click on Create.
2. Click on the Get Started link inside the Try Compute Engine box.



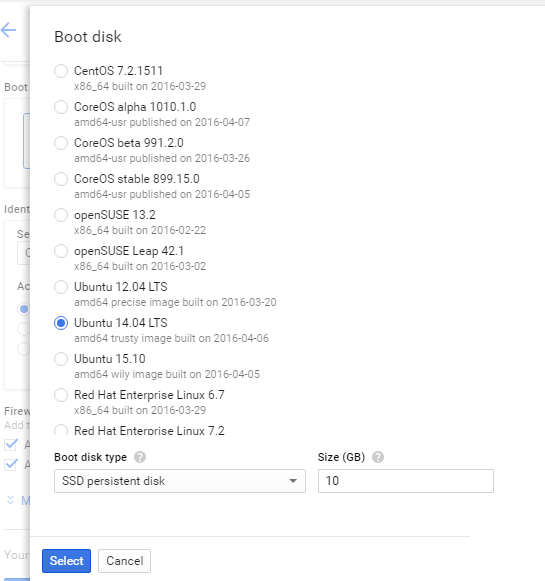
1. Click on Create Instance.

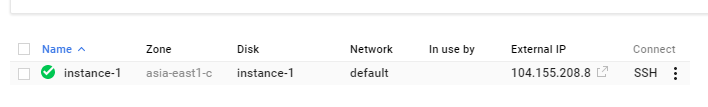


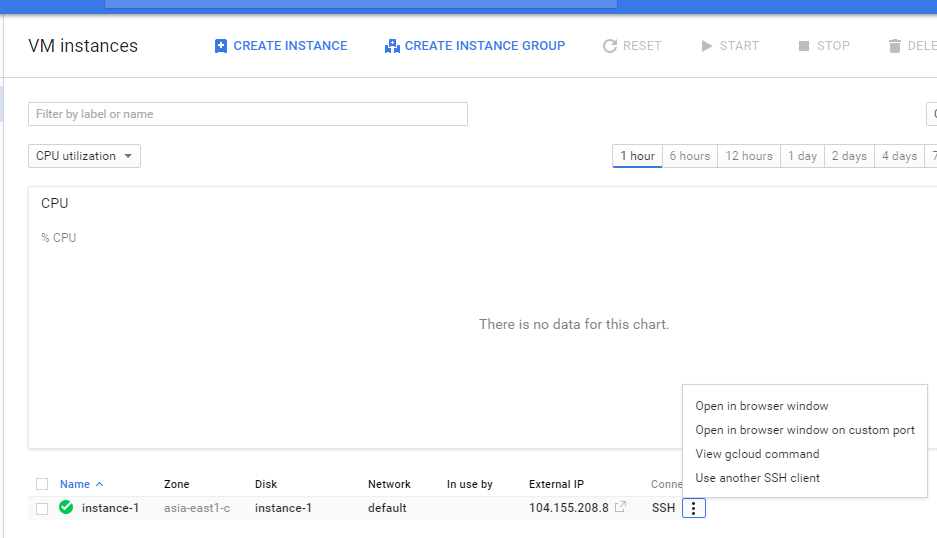
1. Enter Instance details such as name. Under Firewall, check the boxes to allow HTTP and HTTPS access. For my VM, I chose 1 vCPU and 4 GB of memory (you can alter this value by clicking on customize and moving the slider). Check the corresponding price on the right. You have $300 to spend for a period of 2 months. Ensure your configurations are within this limit.



1. Under Boot disk, choose Ubuntu 14.04 LTS



1. Leave other details as is. Click on Create at the end of the screen to create your virtual-machine/instance with Ubuntu OS. Wait while the instance gets created (bottom of screen).
2. To log in to your instance, click on the 3 dots next to SSH under Connect (bottom right of the screen). Choose Open in browser window.



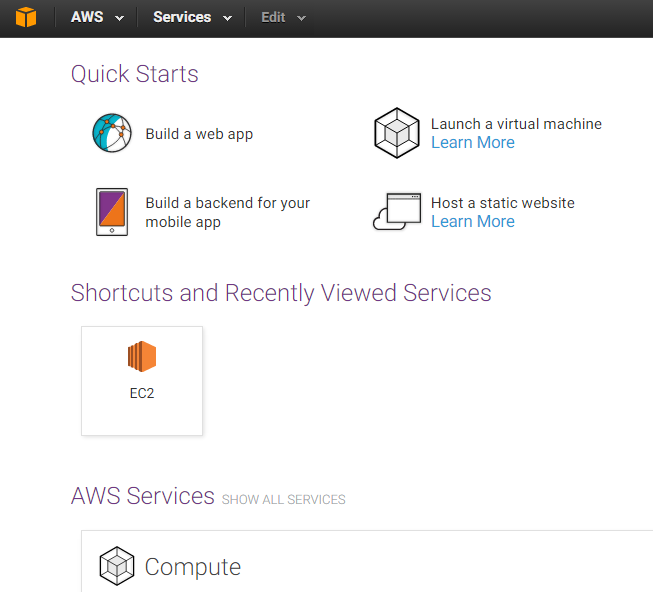
1. Once the browser opens, you will be able to directly run commands (from Step 6) mentioned in the below file:

[**big-data-mapreduce-course**](https://github.com/mahmoudparsian/big-data-mapreduce-course)/[spark](https://github.com/mahmoudparsian/big-data-mapreduce-course/tree/master/spark)/**Install\_Spark\_Pre\_Built\_Version\_by\_Bhushan\_Kumar\_Kothari.pdf**

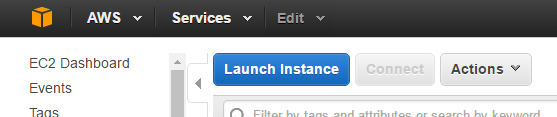
## - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

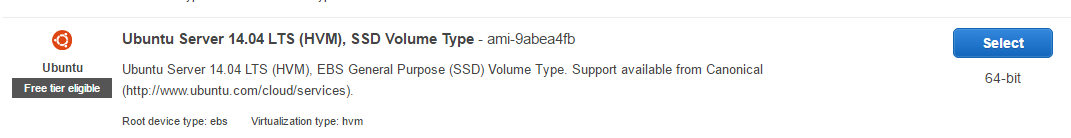
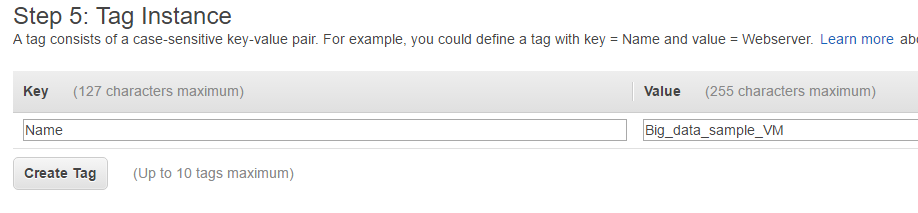
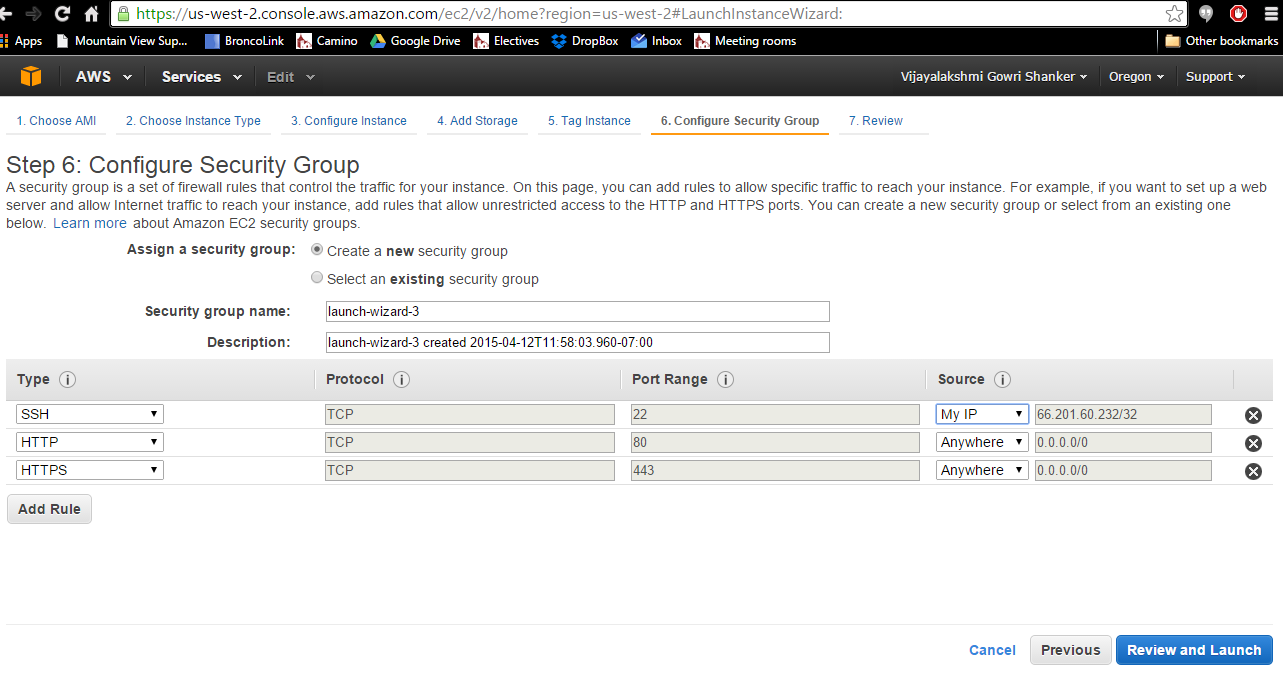
## ALTERNATIVE 2: Installing Ubuntu and Spark on Amazon’s AWS Virtual Machine

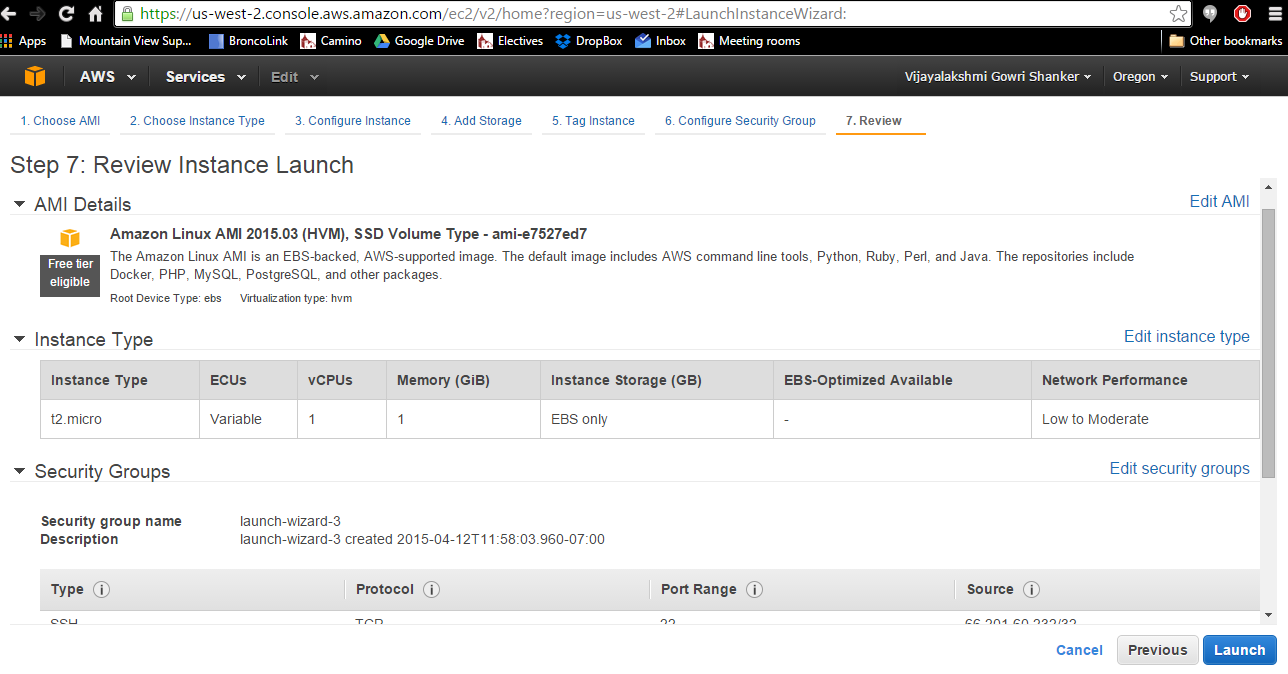
1. **Go to** <http://aws.amazon.com/> **and click on Sign in to the Console.**
2. **If you have an existing account with Amazon, you can use the same credentials to log in. Else, create a new account along with your credit card details.**
3. **Click on EC2.**



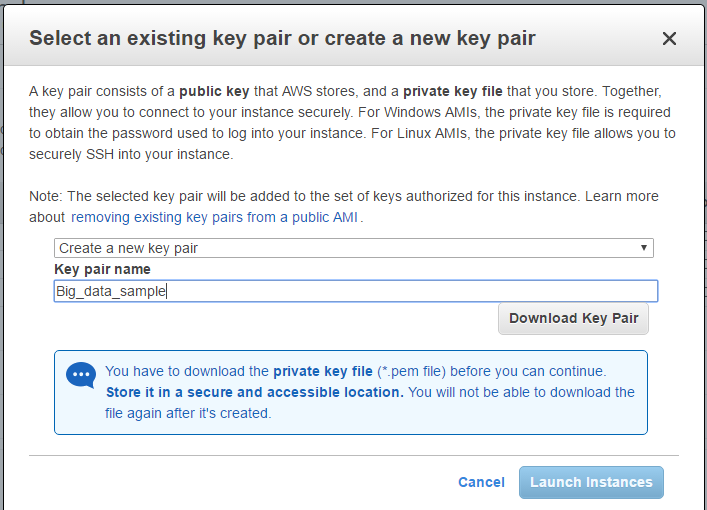
1. Click on Instances on the left panel. Click on Launch Instance on the top left of the screen.



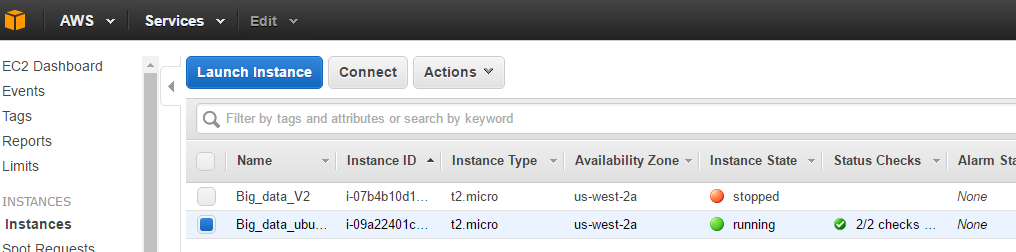
1. Scroll down till you find the Ubuntu image. Alternatively, you can use any Linux image. I chose Ubuntu for convenience. Below Ubuntu, you would be able to see a label: Free tier eligible. Ensure you make similar selections going forward.
2. T2.micro is the size of your instance. Only this is free, so select this and click on Next.
3. Leave contents of Step3: Configure Instance Details page as is. Click on Next.
4. The storage available is 8GB. Do not make any changes and click on Next.
5. You can enter any data for Value. I entered “Big\_data\_sample\_VM”
6. Under security groups, click on Add Rule and add HTTP and HTTPS one after the other. For both these rules, set the source to Anywhere.
7. We can review our selections on this page and can change if necessary. Click on ‘Launch’



1. Click on Launch. In this window, you should create a key-pair.pem file if you are creating an AWS VM for the first time. For subsequent creations, you can either use this key-pair or create a new pair. Choose create a new pair and enter a name for the file. Click on Download Key Pair. **PLEASE** remember where you save this file.



1. Once downloaded, agree to the terms and conditions and click on Launch Instances. Please wait while the instance gets launched.



1. If you want to stop this instance, select that instance, go to Actions -> Instance state -> Stop. Follow the same step for starting the instance. It is advisable to stop your instance once you have executed your programs and derived your output as you have an upper limit on the number of hours of free usage. You can always start it again when you have to run PySpark again.
2. **WINDOWS users**: To connect to the EC2 instance (to start your virtual machine), follow this link: <https://www.youtube.com/watch?v=8Dsq4MeVh8M>

**MAC users** can directly connect from their terminal using the ssh command:

**ssh -i "<name of .pem file>.pem" ubuntu@ec2-54-201-154-64.us-west-2.compute.amazonaws.com**

The above ssh command is specific to my system To get your version, select the instance and click on connect on top of the screen and select “ A standalone SSH client”.

1. Once you have connected to your instance, follow instructions in the below document (from step 6) to start working with PySpark:

[**big-data-mapreduce-course**](https://github.com/mahmoudparsian/big-data-mapreduce-course)/[spark](https://github.com/mahmoudparsian/big-data-mapreduce-course/tree/master/spark)/**Install\_Spark\_Pre\_Built\_Version\_by\_Bhushan\_Kumar\_Kothari.pdf**