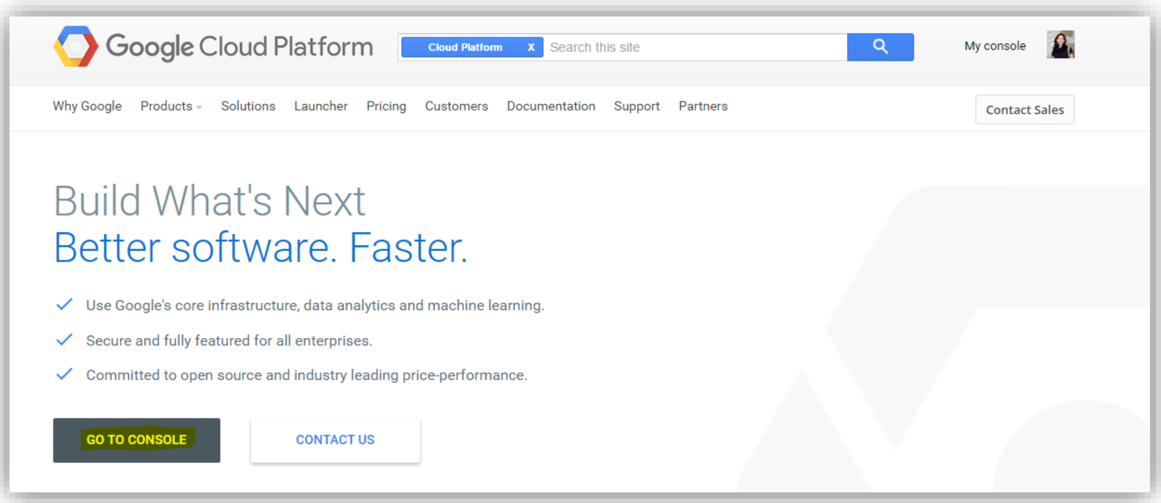
ALTERNATIVE SOLUTIONS FOR INSTALLING AND RUNNING SPARK

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.

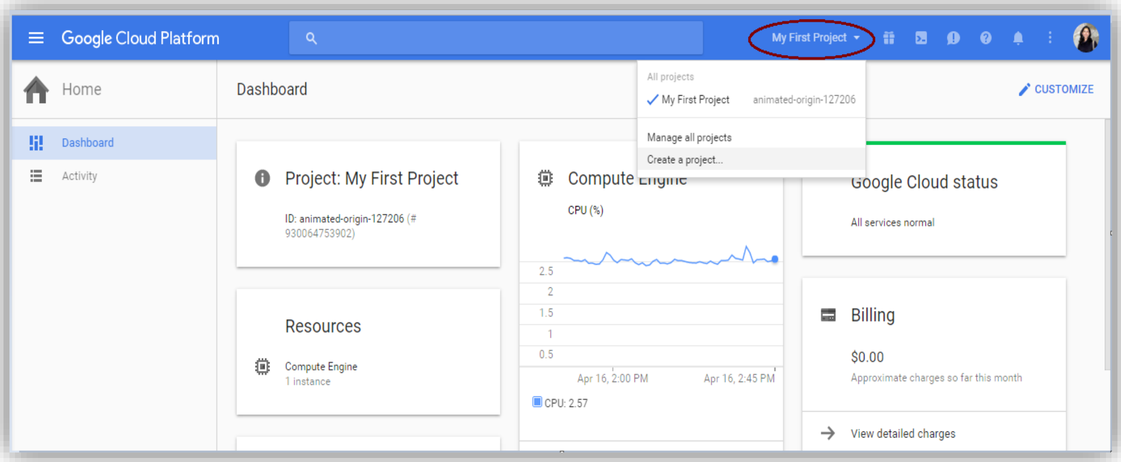


Build What's Next
Better software. Faster.

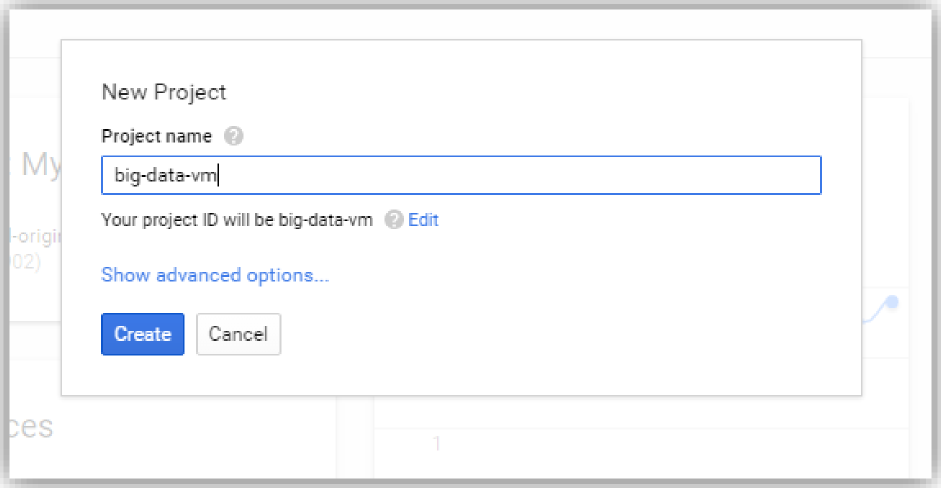
✓ Use Google's core infrastructure, data analytics and machine learning.
✓ Secure and fully featured for all enterprises.
✓ Committed to open source and industry leading price-performance.

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.



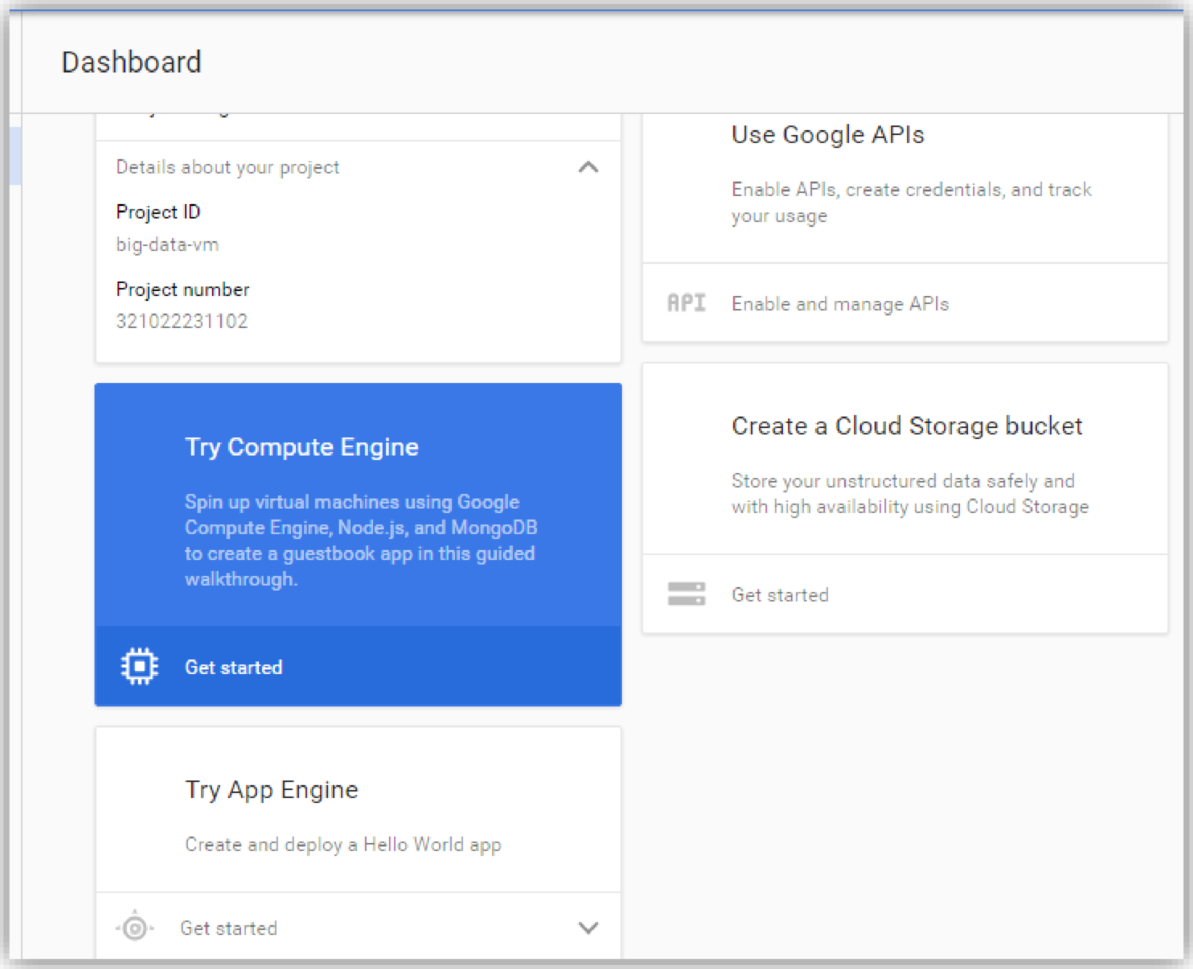
5. Give a suitable name for your project (example: big-data-vm) and click on Create.



Project name: big-data-vm

Your project ID will be big-data-vm Edit

6. Click on the Get Started link inside the Try Compute Engine box.



Dashboard

Details about your project

Project ID

big-data-vm

Project number

321022231102

Try Compute Engine

Spin up virtual machines using Google Compute Engine, Node.js, and MongoDB to create a guestbook app in this guided walkthrough.

Try App Engine

Create and deploy a Hello World app

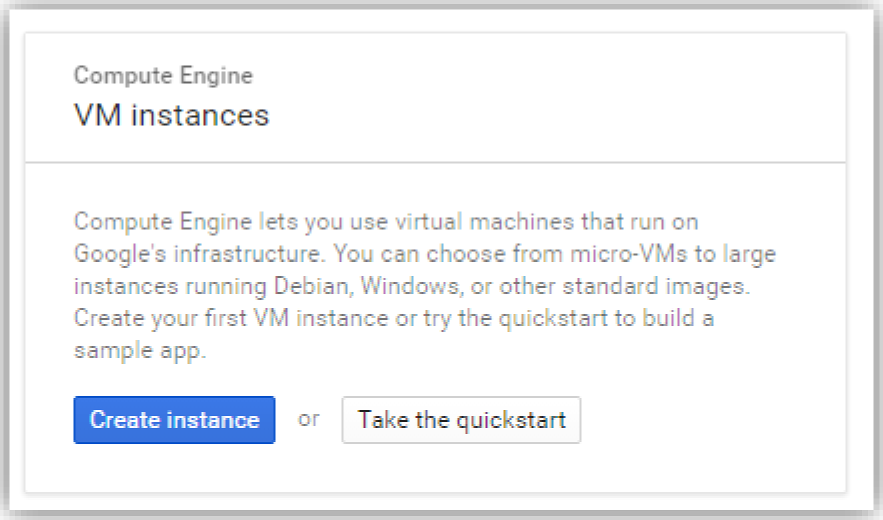
Use Google APIs

Enable APIs, create credentials, and track your usage

Create a Cloud Storage bucket

Store your unstructured data safely and with high availability using Cloud Storage

7. Click on Create Instance.

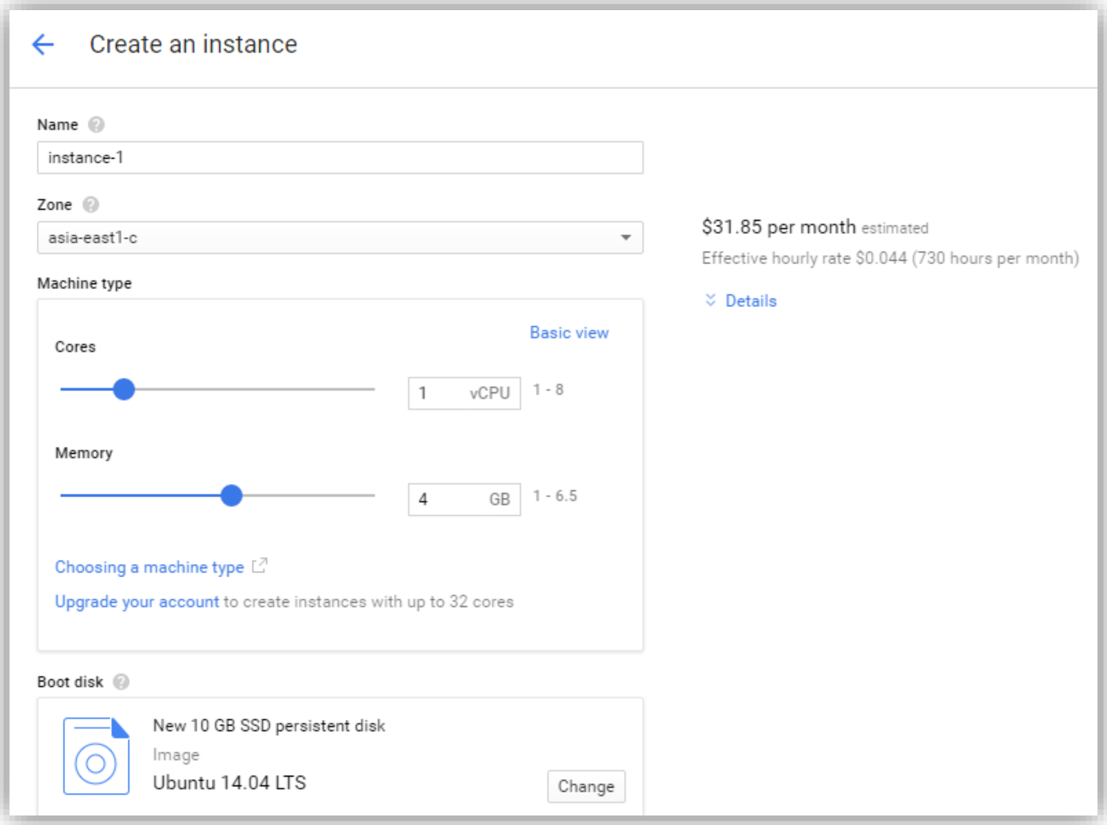


Compute Engine

VM instances

Compute Engine lets you use virtual machines that run on Google's infrastructure. You can choose from micro-VMs to large instances running Debian, Windows, or other standard images. Create your first VM instance or try the quickstart to build a sample app.

8. 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.



Create an instance

Name: instance-1

Zone: asia-east1-c

Machine type

Cores: 1 vCPU (1 - 8)

Memory: 4 GB (1 - 6.5)

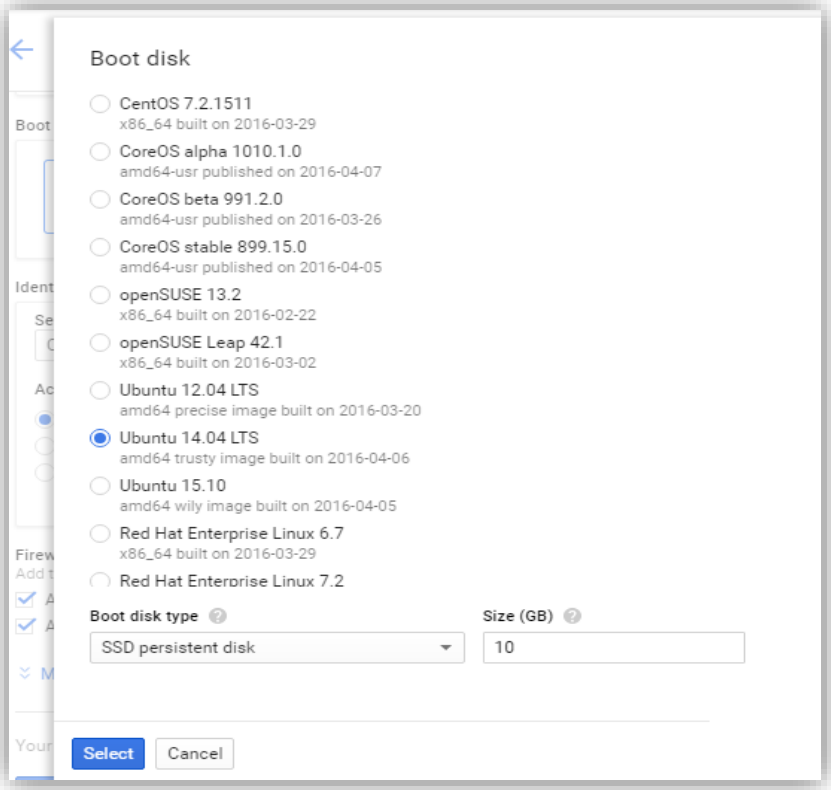
Upgrade your account to create instances with up to 32 cores

Boot disk: New 10 GB SSD persistent disk, Image Ubuntu 14.04 LTS

$31.85 per month estimated

Effective hourly rate $0.044 (730 hours per month)

9. Under Boot disk, choose Ubuntu 14.04 LTS



Boot disk

○ CentOS 7.2.1511
x86_64 built on 2016-03-29

○ CoreOS alpha 1010.1.0
amd64-usr published on 2016-04-07

○ CoreOS beta 991.2.0
amd64-usr published on 2016-03-26

○ CoreOS stable 899.15.0
amd64-usr published on 2016-04-05

○ openSUSE 13.2
x86_64 built on 2016-02-22

○ openSUSE Leap 42.1
x86_64 built on 2016-03-02

○ Ubuntu 12.04 LTS
amd64 precise image built on 2016-03-20

● Ubuntu 14.04 LTS
amd64 trusty image built on 2016-04-06

○ Ubuntu 15.10
amd64 wily image built on 2016-04-05

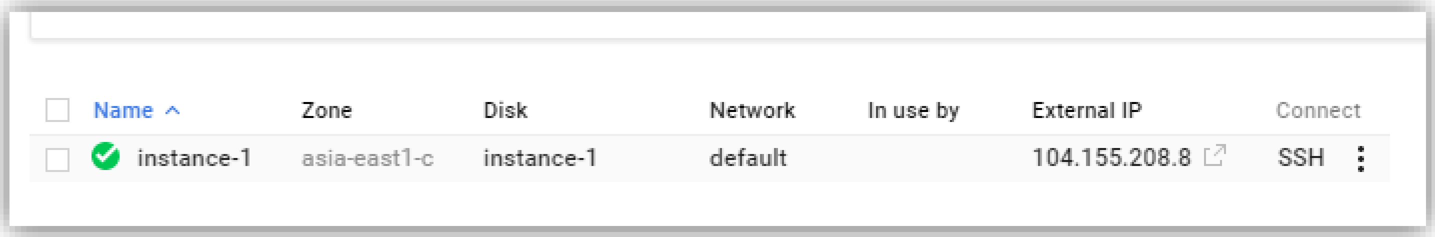
○ Red Hat Enterprise Linux 6.7
x86_64 built on 2016-03-29

○ Red Hat Enterprise Linux 7.2

Boot disk type: SSD persistent disk

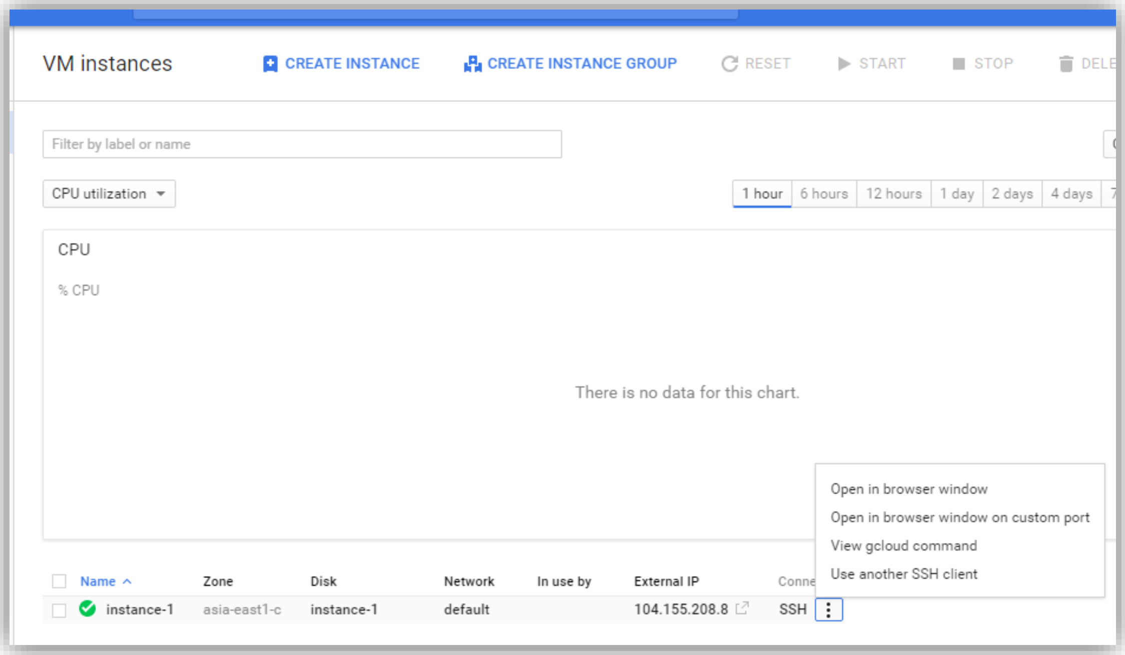
Size (GB): 10

10. 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).



□	Name ^	Zone	Disk	Network	In use by	External IP	Connect
□	✓ instance-1	asia-east1-c	instance-1	default		104.155.208.8	SSH ⋮

11. 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.



There is no data for this chart.

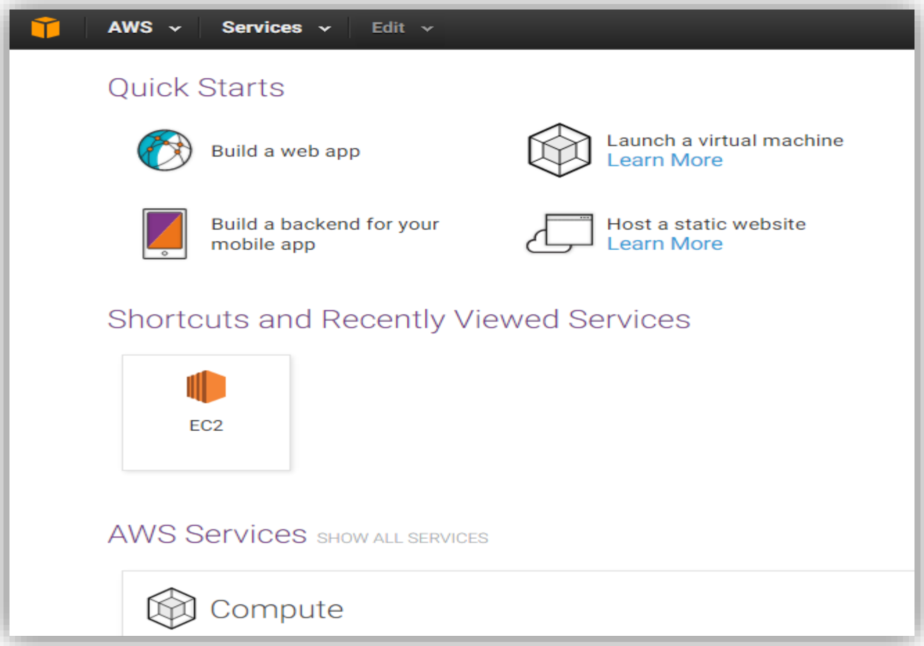
Name	Zone	Disk	Network	In use by	External IP	Conne
instance-1	asia-east1-c	instance-1	default		104.155.208.8	SSH

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

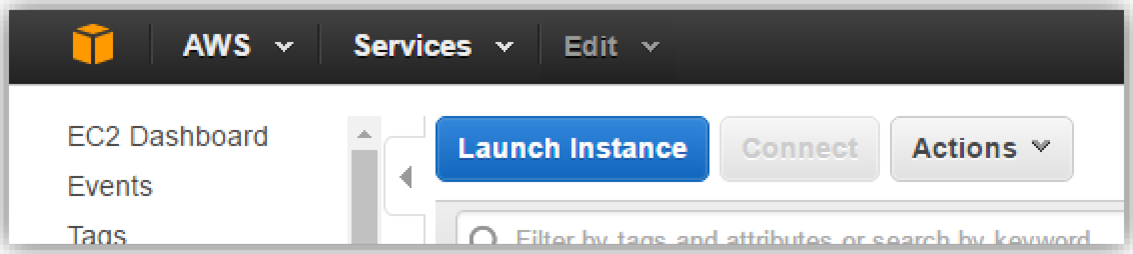
    big-data-mapreduce-course/spark/**Install_Spark_Pre_Built_Version_by_Bhushan_Kumar_Kothari.pdf**

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

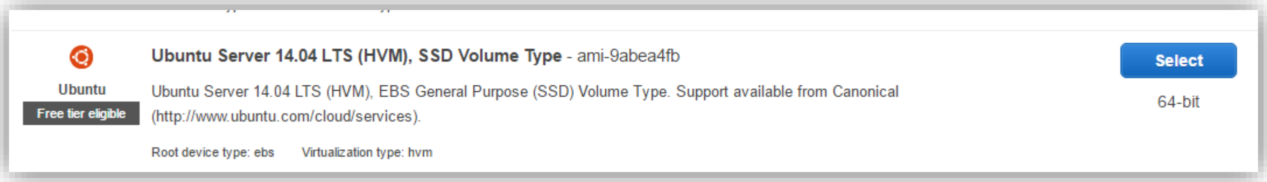
1. Go to http://aws.amazon.com/ and click on Sign in to the Console.
2. If you have an existing account with Amazon, you can use the same credentials to log in. Else, create a new account along with your credit card details.
3. Click on EC2.



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



5. 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.



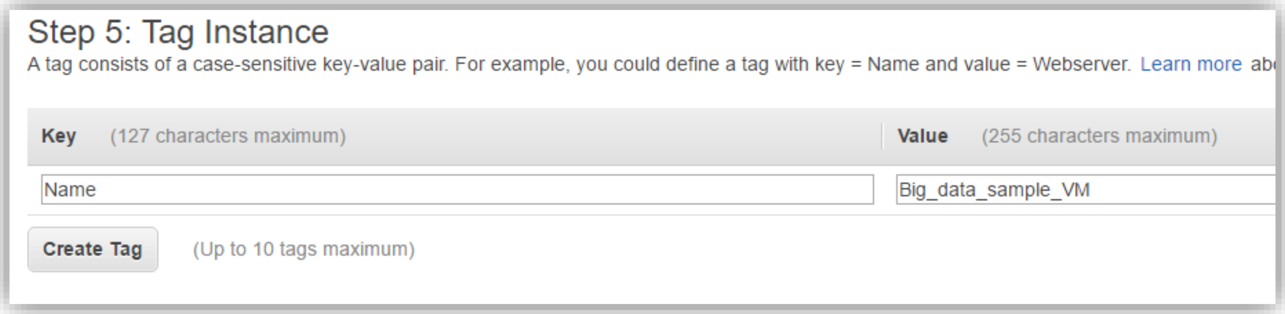
Ubuntu Server 14.04 LTS (HVM), SSD Volume Type - ami-9abea4fb

Ubuntu Server 14.04 LTS (HVM), EBS General Purpose (SSD) Volume Type. Support available from Canonical (http://www.ubuntu.com/cloud/services).

Root device type: ebs Virtualization type: hvm

6. T2.micro is the size of your instance. Only this is free, so select this and click on Next.
7. Leave contents of Step3: Configure Instance Details page as is. Click on Next.
8. The storage available is 8GB. Do not make any changes and click on Next.

9. You can enter any data for Value. I entered “Big_data_sample_VM”

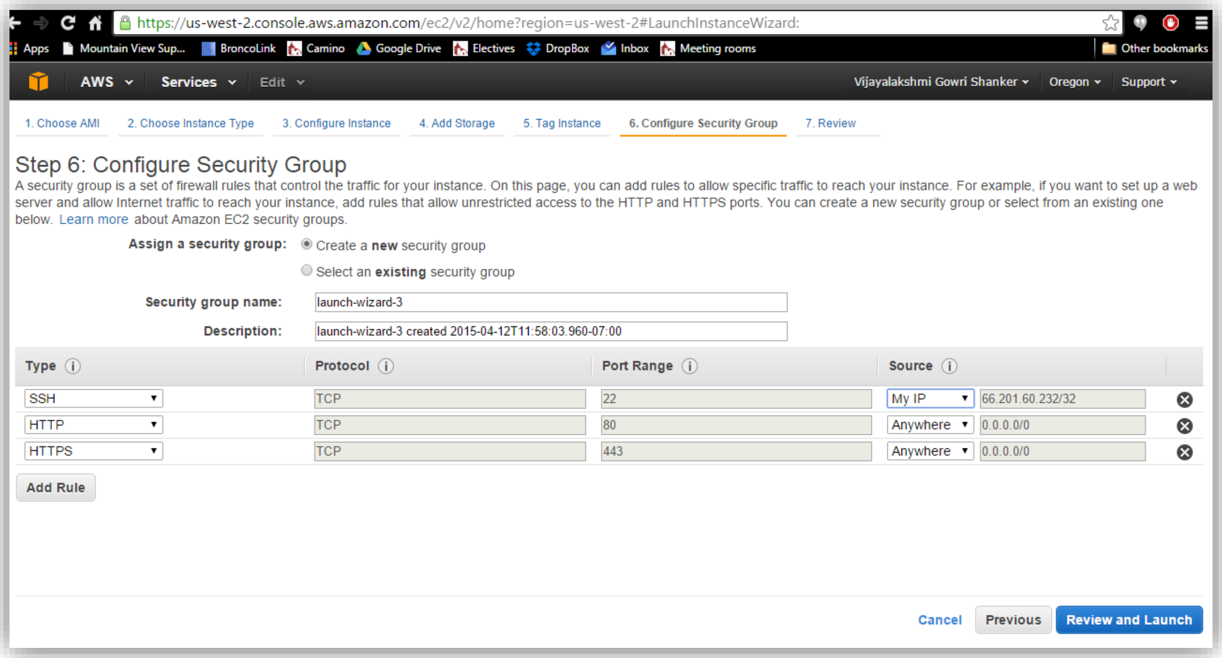


Step 5: Tag Instance

A tag consists of a case-sensitive key-value pair. For example, you could define a tag with key = Name and value = Webserver. Learn more ab

Key (127 characters maximum)	Value (255 characters maximum)
Name	Big_data_sample_VM

10. 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.



Step 6: Configure Security Group

A security group is a set of firewall rules that control the traffic for your instance. On this page, you can add rules to allow specific traffic to reach your instance. For example, if you want to set up a web server and allow Internet traffic to reach your instance, add rules that allow unrestricted access to the HTTP and HTTPS ports. You can create a new security group or select from an existing one below. Learn more about Amazon EC2 security groups.

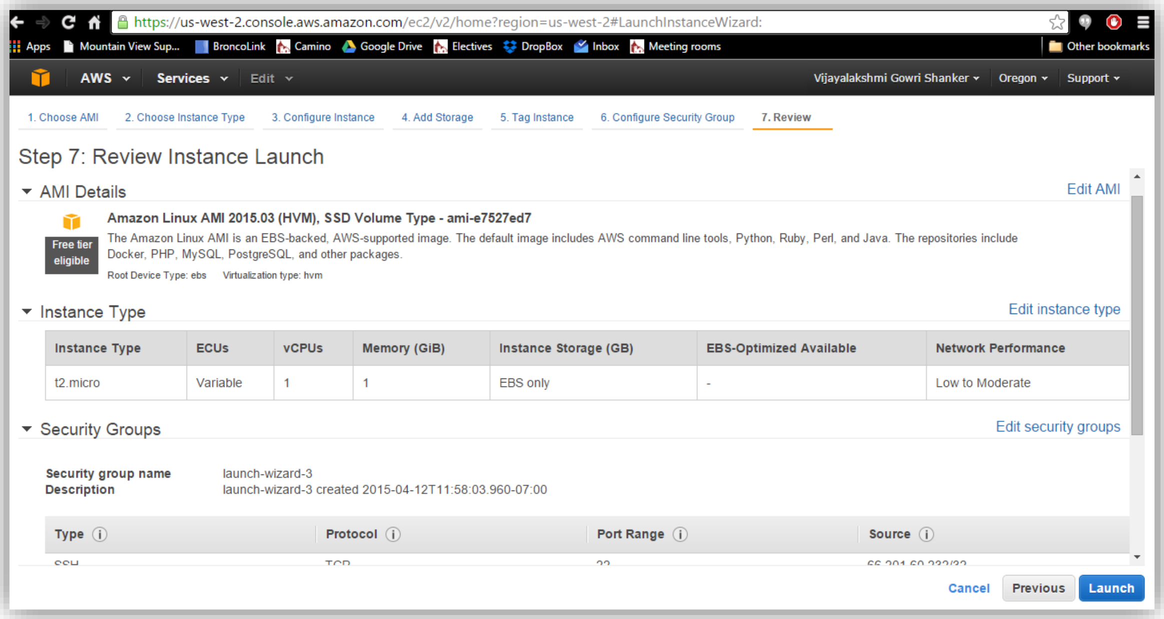
Assign a security group: ● Create a **new** security group
○ Select an **existing** security group

Security group name: launch-wizard-3

Description: launch-wizard-3 created 2015-04-12T11:58:03.960-07:00

Type	Protocol	Port Range	Source	
SSH	TCP	22	My IP	66.201.60.232/32
HTTP	TCP	80	Anywhere	0.0.0.0/0
HTTPS	TCP	443	Anywhere	0.0.0.0/0

11. We can review our selections on this page and can change if necessary. Click on ‘Launch’



Step 7: Review Instance Launch

AMI Details

Free tier eligible

Amazon Linux AMI 2015.03 (HVM), SSD Volume Type - ami-e7527ed7

The Amazon Linux AMI is an EBS-backed, AWS-supported image. The default image includes AWS command line tools, Python, Ruby, Perl, and Java. The repositories include Docker, PHP, MySQL, PostgreSQL, and other packages.

Root Device Type: ebs Virtualization type: hvm

Instance Type

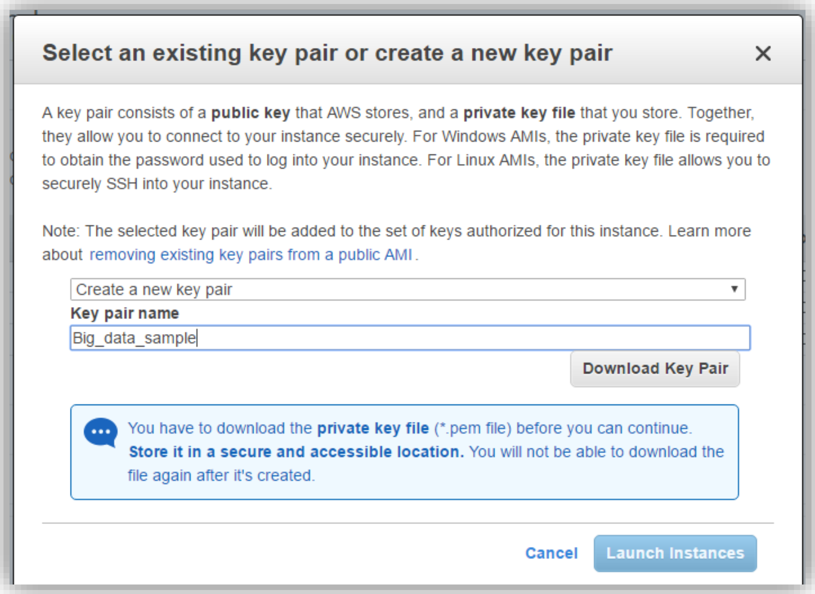
Instance Type	ECUs	vCPUs	Memory (GiB)	Instance Storage (GB)	EBS-Optimized Available	Network Performance
t2.micro	Variable	1	1	EBS only	-	Low to Moderate

Security Groups

Security group name launch-wizard-3
Description launch-wizard-3 created 2015-04-12T11:58:03.960-07:00

Type	Protocol	Port Range	Source
SSH	TCP	22	66.201.60.232/32

12. 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.



Select an existing key pair or create a new key pair

A key pair consists of a **public key** that AWS stores, and a **private key file** that you store. Together, they allow you to connect to your instance securely. For Windows AMIs, the private key file is required to obtain the password used to log into your instance. For Linux AMIs, the private key file allows you to securely SSH into your instance.

Note: The selected key pair will be added to the set of keys authorized for this instance. Learn more about removing existing key pairs from a public AMI .

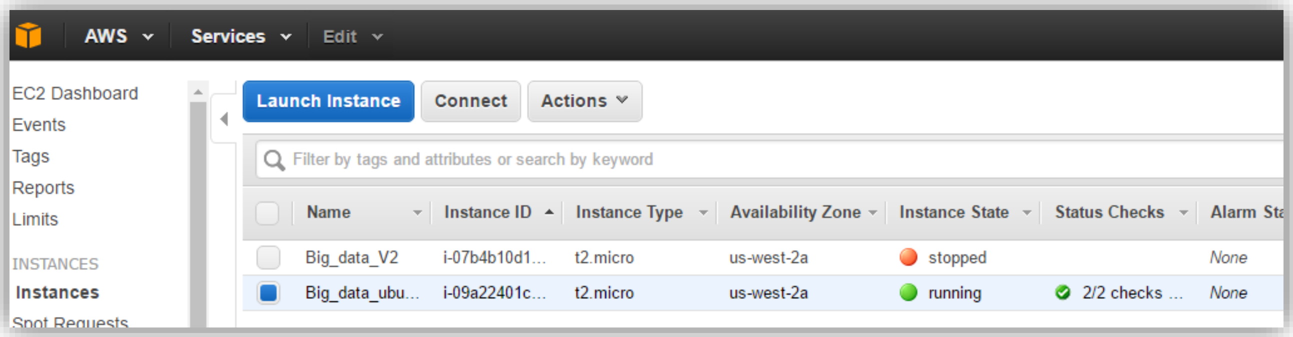
Create a new key pair

Key pair name

Big_data_sample

You have to download the **private key file** (*.pem file) before you can continue. **Store it in a secure and accessible location.** You will not be able to download the file again after it's created.

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



☐	Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks	Alarm Sta
☐	Big_data_V2	i-07b4b10d1...	t2.micro	us-west-2a	● stopped		None
☑	Big_data_ubu...	i-09a22401c...	t2.micro	us-west-2a	● running	✓ 2/2 checks ...	None

14. 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.

15. **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”.

16. 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/spark/**Install_Spark_Pre_Built_Version_by_Bhushan_Kumar_Kothari.pdf**