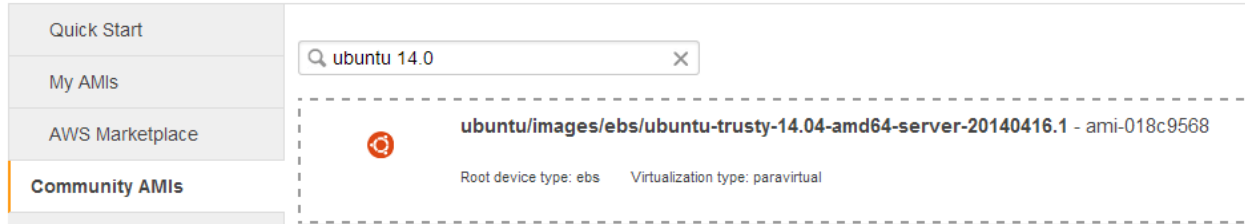


Multicore Kmeans amazon instance run documentation

Step 1: Logon to Amazon AWS and under Services select 'Ec2'. Click on Launch Instance button.

Step 2: Under the Quick Start column on the left, click on Community AMIs

Step 3: Search for "ubuntu 14.0", select the first AMI "ubuntu/images/ebs/ubuntu-trusty-14.04-amd64-server-20140416.1 - ami-018c9568". Click on 'Select' on the right side of the instance.



Step 4: Select the 'Compute Optimized', 'c3.2xlarge' instance and click on 'Next: Configure Instance Details'; at the bottom right.

<input type="checkbox"/>	Compute optimized	c3.xlarge	14	4	7.5	2 x 40 (SSD)	Yes
<input checked="" type="checkbox"/>	Compute optimized	c3.2xlarge	28	8	15	2 x 80 (SSD)	Yes
<input type="checkbox"/>	Compute optimized	c3.4xlarge	55	16	30	2 x 160 (SSD)	Yes

Step 5: Keep everything same and click on 'Next: Add Storage'.

Step 6: Increase the size to 20 Gib and change the 'type' from 'Magnetic' to 'General Purpose (SSD)' and click on 'Next: Tag Instance'.

Type ⓘ	Device ⓘ	Snapshot ⓘ	Size (GiB) ⓘ	Volume Type ⓘ	IOPS ⓘ	Delete on Termination ⓘ
Root	/dev/sda1	snap-7ac25ca6	<input type="text" value="20"/>	General Purpose (SSD) ▼	60 / 3000	<input checked="" type="checkbox"/>
Instance Store 0 ▼	/dev/sdb ▼	N/A	N/A	N/A	24 / 3000	N/A

Step 7: Name the instance as 'MulticoreInstance' and click on 'Next: Configure Security Group'.

Key (127 characters maximum)	Value (255 characters maximum)
<input type="text" value="Name"/>	<input type="text" value="MulticoreInstance"/>

(Up to 10 tags maximum)

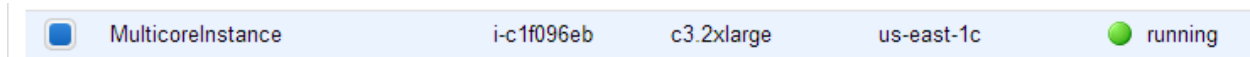
Step 8: Next comes the 'Configure Security Group', keep the default settings and click "Next:

Review and Launch”.

Step 9: Just check that what we configured is showing up. Click on “launch”.

Step 10: In the 'Select existing Key pair or create a new key pair' dialog box, select create new key pair and name it. We gave the name 'dmkd_cpu'. This key pair file will be used to login to the instance. Click on ‘Download Key Pair’ and click on Launch instance.

Step 11: If everything went well Launch Status page will show up with “Your instance is now launching” statement. Click on “View Instance” button on bottom right. You could see you instance in the running instances by the name '**MulticoreInstance**'.



Step 12: When you select the instance, Instance description shows up at the bottom. Copy the 'Public DNS' value, it will be something like 'ec2-.....-amazonaws.com'.

Windows Login

Step 1: Open '[Puttygen](#)' and in the dialogbox click on ‘Load’ and select the keypair file which you downloaded in Step 10 “dmkd_cpu.pem”. In the search dialog box, select ‘All files’ at the bottom right which will show the ‘.pem’ file. Click on open and click ‘OK’ for successful import notice. Click on ‘Save private key’ at the bottom right and click ‘yes’ to the warning. Save the file with the same name and without the ‘.pem’ extension, Putty will automatically add the .ppk extension to the newly created file.

Step 2: Open [Putty](#) and in the ‘Host name’ type ‘ubuntu@<DNS value which you copied in Step 12>’. Like ‘ubuntu@ec2-.....-amazonaws.com’.

Step 3: Under the connection category in the left panel, select the ‘+’ near ‘SSH’ and click on ‘Auth’. Browse for the ‘.ppk’ which we created in Step 12 and click open. If everything goes well, it will connect to the amazon instance.

Copying Code

Step 1: Open '[WinSCP](#)' and in the dialog box for ‘Host name’, paste the DNS name which you copied in Step 12 of instance creation. Under User name, type “ubuntu” and click on ‘Advanced’ under Password text box.

Step 2: In the ‘Advanced Site Settings’ pop up, click on ‘Authentication’ under ‘SSH’ in the left column. Browse for the Private Key file ‘.ppk’ which you created in Step 1 of ‘Windows Login’.

Step 3: Click on ‘Login’ in the main window and it will connect you the amazon AWS instance.

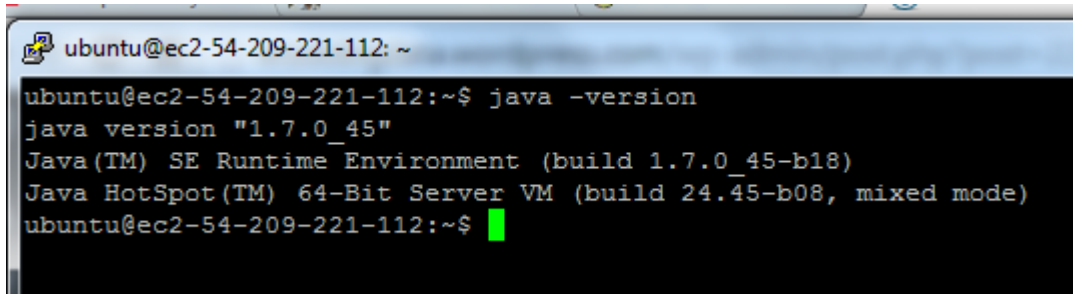
Step 4: Browse to the Multicore_Kmeans code in the left window and copy the folder on the right side (on instance, /home/ubuntu). Close WinSCP, once done.

Running the Code

Step 1: Follow Step 2 of Windows Login and connect to the instance.

Step 2: Install Java. Add following PPA and install the latest Oracle Java (JDK) 7 in Ubuntu

- a) Type '\$ sudo add-apt-repository ppa:webupd8team/java'
- b) Type '\$ sudo apt-get update && sudo apt-get install oracle-jdk7-installer'
- c) Check if Ubuntu uses JDK 7 by typing 'java -version'. It will show something like this



```
ubuntu@ec2-54-209-221-112: ~  
ubuntu@ec2-54-209-221-112:~$ java -version  
java version "1.7.0_45"  
Java(TM) SE Runtime Environment (build 1.7.0_45-b18)  
Java HotSpot(TM) 64-Bit Server VM (build 24.45-b08, mixed mode)  
ubuntu@ec2-54-209-221-112:~$
```

Step 3: Browse to the Multicore_Kmeans directory and check if we have the following:

- a) 'Input' directory
- b) JAR file named 'ProcessCorpus.jar'
- c) JAR file named 'GetCentroids.jar'
- d) JAR file named 'KMeans.jar'

Step 4: Type '**java -jar ProcessCorpus.jar**'. When prompted, answer:

```
Enter the directory where the corpus is located: Input  
Enter the name of the file to write the result to: vectors  
Enter the max number of docs to use in each subdirectory: 100  
How many of those words do you want to use...? 10000
```

This will create a file called "vectors" that has 20 * 100 lines, each of which is a vectorized representation of a document. The dictionary size that is used is 10000 words

Step 5: Now type '**java -jar GetCentroids.jar**'. When prompted, answer:

```
Enter the data file to select the clusters from: vectors  
Enter the name of the file to write the result to: clusters  
Enter the number of clusters to select: 20
```

This will create a file called "clusters" that has 20 lines, each of which describes a cluster centroids. This initial set of cluster centroids is simply randomly sampled from the "vectors" file.

Step 6: Run the kmeans jar by typing

```
$ Java -jar Kmeans.jar
```

Step 7: When done, check the 'clusters' file to get the cluster distribution.

Cleanup (Important)

Step 1: Logon to Amazon AWS and under Services select 'Ec2'.

Step 2: Under the 'Instances' tab in the left column; click on 'Instances'.

Step 3: Locate your instance (here MulticoreInstance) and select it. On the top locate 'Actions' drop down button and click 'Stop' to stop instance. You can start it and connect to the same settings whenever you want. If you terminate it, you have to create a new instance all together.