Running Multicore K-Nearest Neighbor on amazon EC2

Pseudo Code

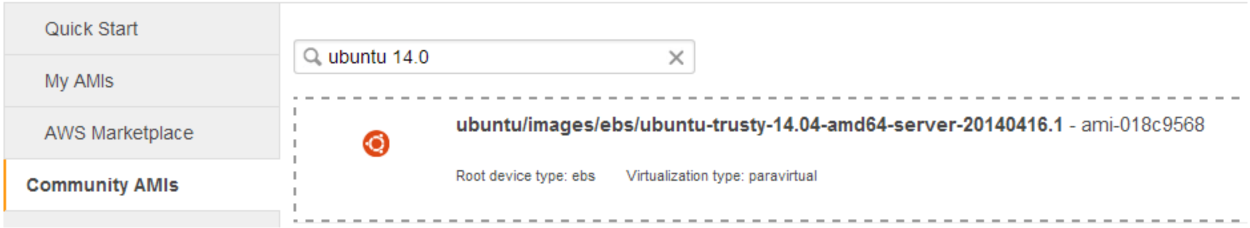
```
Input: Reference points R, Query points Q
Step 1: Compute the distance of each point in Q with points in R.
Step 2: Create a Distance matrix $m \times n$ where m are the query points and n are the
        distance points.
Step 3: Sort each row in the matrix
Step 4: Select the K nearest neighbors for each query point
Output: Query point with K nearest neighbors
```

How to run the code

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.

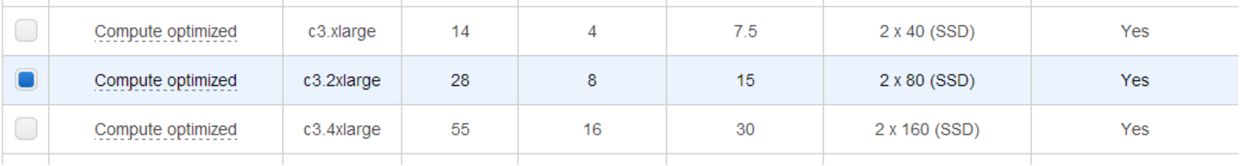


ubuntu 14.0

ubuntu/images/ebs/ubuntu-trusty-14.04-amd64-server-20140416.1 - ami-018c9568

Root device type: ebs   Virtualization type: paravirtual

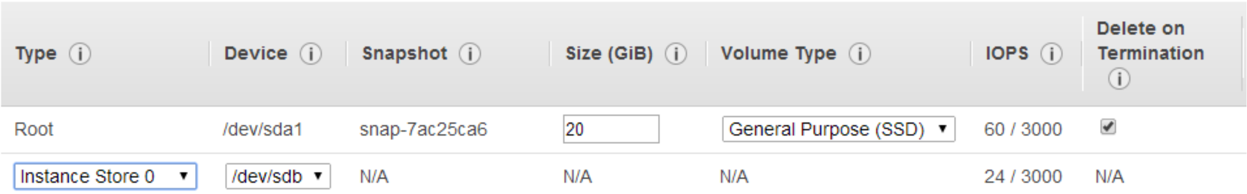
Step 4: Select the 'Compute Optimized', 'c3.2xlarge' instance and click on 'Next: Configure Instance Details'; at the bottom right.



☐	Compute optimized	c3.xlarge	14	4	7.5	2 x 40 (SSD)	Yes
☑	Compute optimized	c3.2xlarge	28	8	15	2 x 80 (SSD)	Yes
☐	Compute optimized	c3.4xlarge	55	16	30	2 x 160 (SSD)	Yes

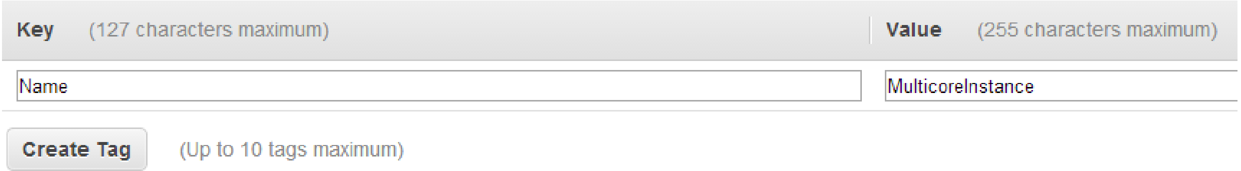
Step 5: Keep everything same and click on 'Next: Add Storgae'.

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



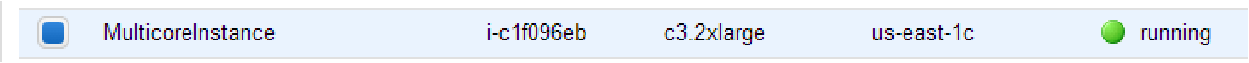
Type ⓘ	Device ⓘ	Snapshot ⓘ	Size (GiB) ⓘ	Volume Type ⓘ	IOPS ⓘ	Delete on Termination ⓘ
Root	/dev/sda1	snap-7ac25ca6	20	General Purpose (SSD) ▾	60 / 3000	☑
Instance Store 0 ▾	/dev/sdb ▾	N/A	N/A	N/A	24 / 3000	N/A

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



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

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**'.



■	MulticoreInstance	i-c1f096eb	c3.2xlarge	us-east-1c	● running

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_KNN code in the left window and copy the folder on the right side (on instance, /home/ubuntu). Close WinSCP, once done.

Running the Code

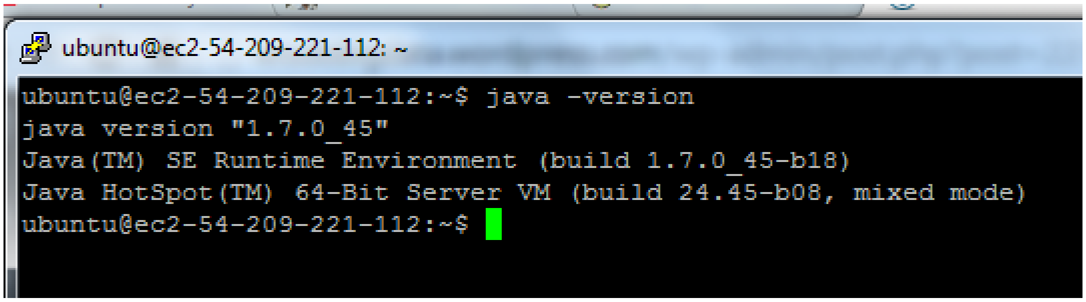
Step 1: Follow Step 2 of Windows Login and connect to the instance.

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_KNN directory and check if we have the following:

a) ‘build’ directory
b) ‘nbproject’ directory
c) ‘src’ directory
d) ‘sampledata’ directory
e) ‘build.xml’ file
f) ‘manifest.mf’ file

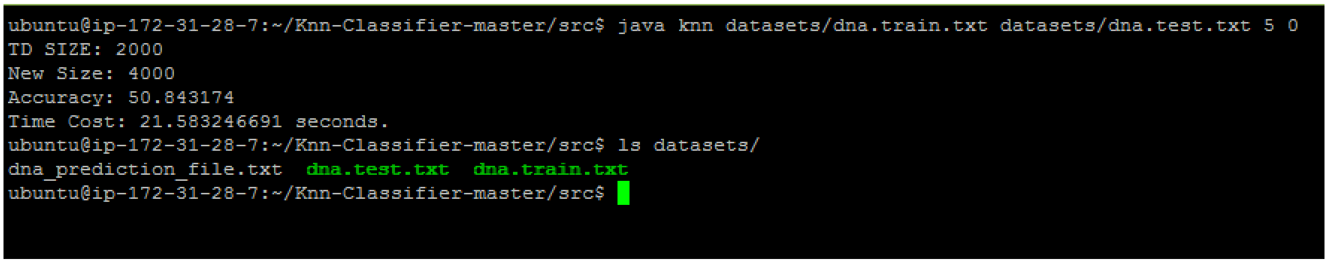
Step 4: Browse to the ‘src’ directory by typing

$ cd src/

Check if we have the directory named ‘Datasets’ there

Now type the following

$ java knn datasets/dna.train.txt datasets/dna.test.txt 5 0



```
ubuntu@ip-172-31-28-7:~/Knn-Classifier-master/src$ java knn datasets/dna.train.txt datasets/dna.test.txt 5 0
TD SIZE: 2000
New Size: 4000
Accuracy: 50.843174
Time Cost: 21.583246691 seconds.
ubuntu@ip-172-31-28-7:~/Knn-Classifier-master/src$ ls datasets/
dna_prediction_file.txt  dna.test.txt  dna.train.txt
ubuntu@ip-172-31-28-7:~/Knn-Classifier-master/src$
```

There are 4 arguments to the above command

```
1. Args[0]-> location of the training data file
2. Args[1]-> location of the test data file
3. Args[2]-> Number of Nearest neighbors(K)
4. Args[3]-> Distance metrics
     a. 0 for Euclidean
     b. 1 for Cosine
```

Step 7: When done, check the ‘datasets/dna_prediction_file.txt’ file to get the classification.

The command line result shows the Accuracy in classifying with the selected distance measure and the time taken to execute the code.

There are many other sample datasets provided under ‘Multicore_KNN/sampledata’ for trying out the algorithm with other data.

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.