Geo-Location Clustering with K-Mean

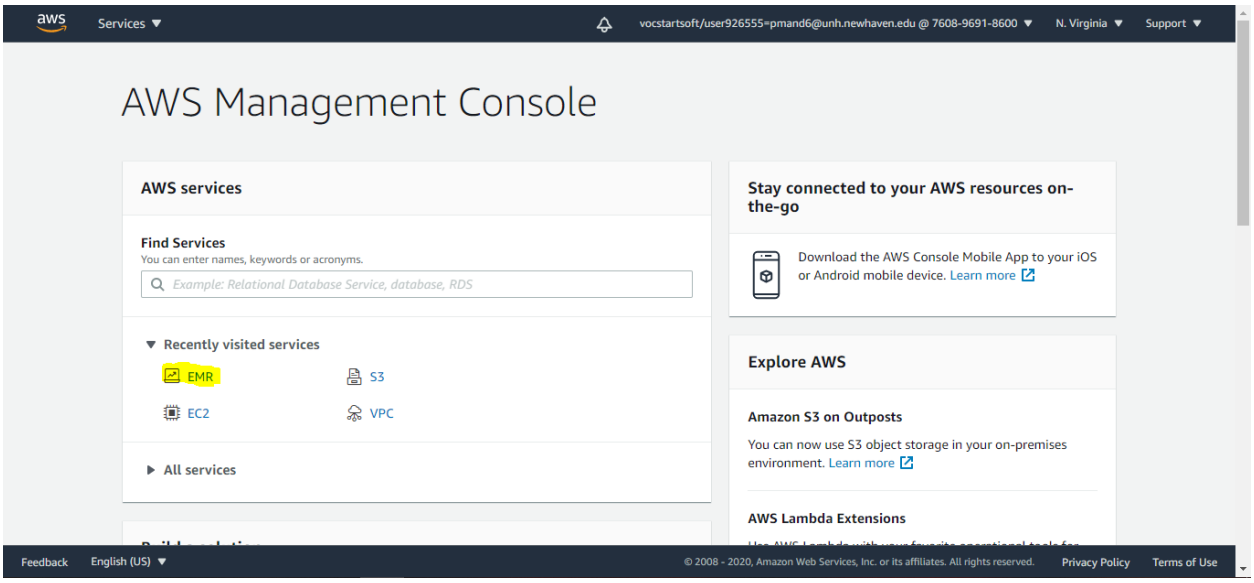
By:
Praveen Mandadi

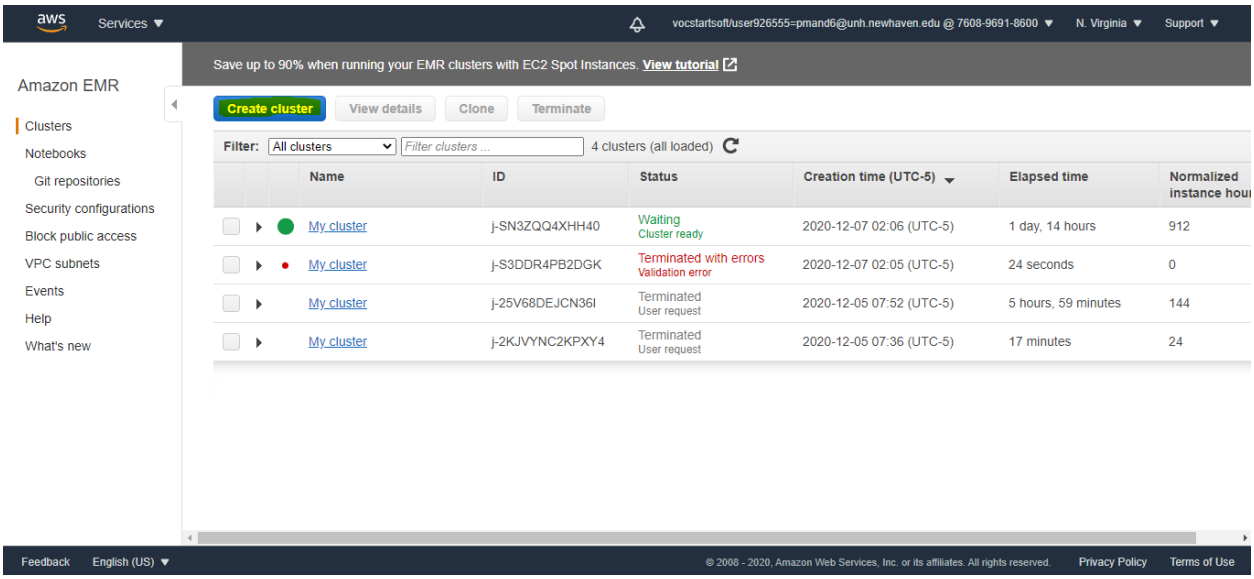
Introduction and Motivation

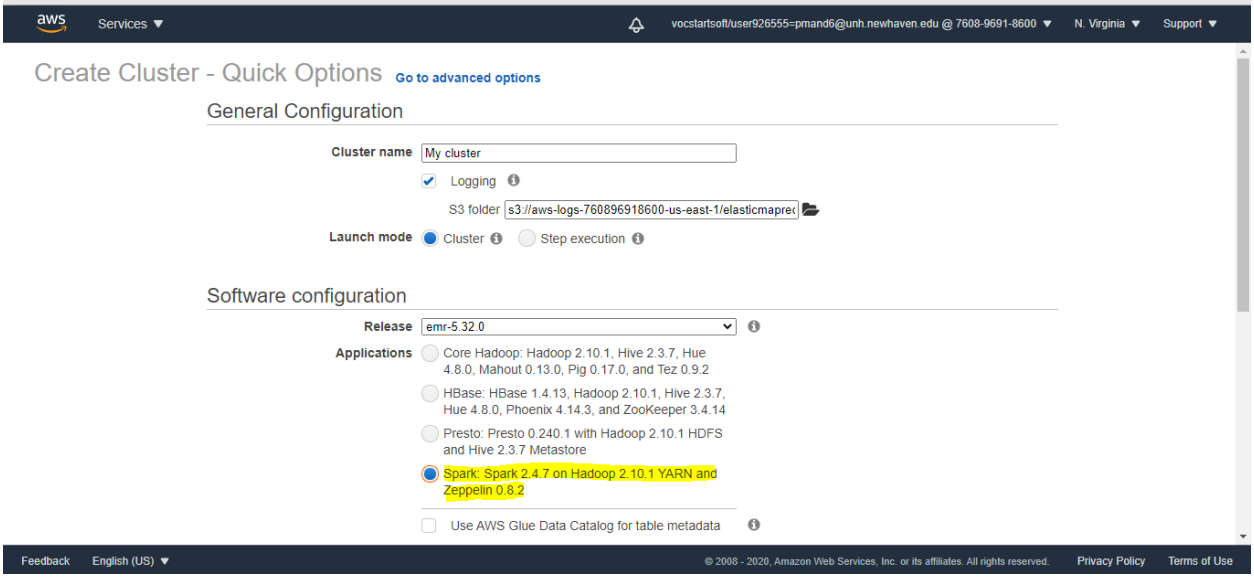
Clustering is a process of grouping a set of data points into clusters so that points that are put within the same cluster are like each other whereas points from different clusters are dissimilar. Clustering has many useful applications for marketing, logistics, and document classification. We clustered geo-location data, which will be naturally visualized.

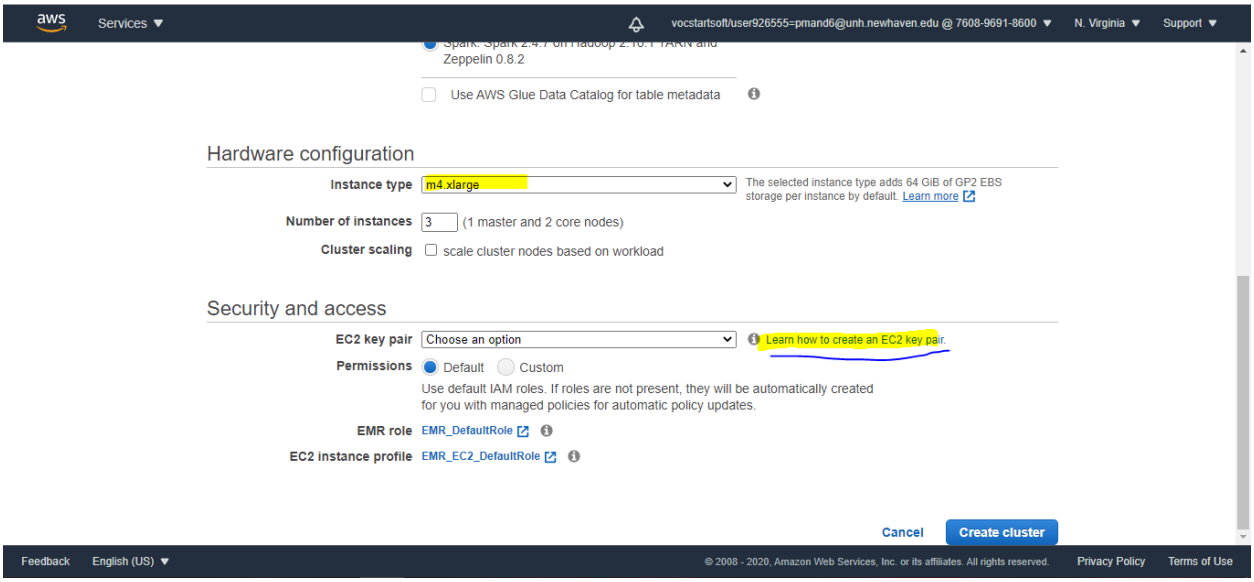
We implemented the k-means algorithm in Spark to solve the clustering issue in a parallel fashion. The algorithm iteratively updates the position of k cluster centroids as a distance-based approach until the shift in the mean of centroids converges to alpha=0.1 km where alpha is conve rged.

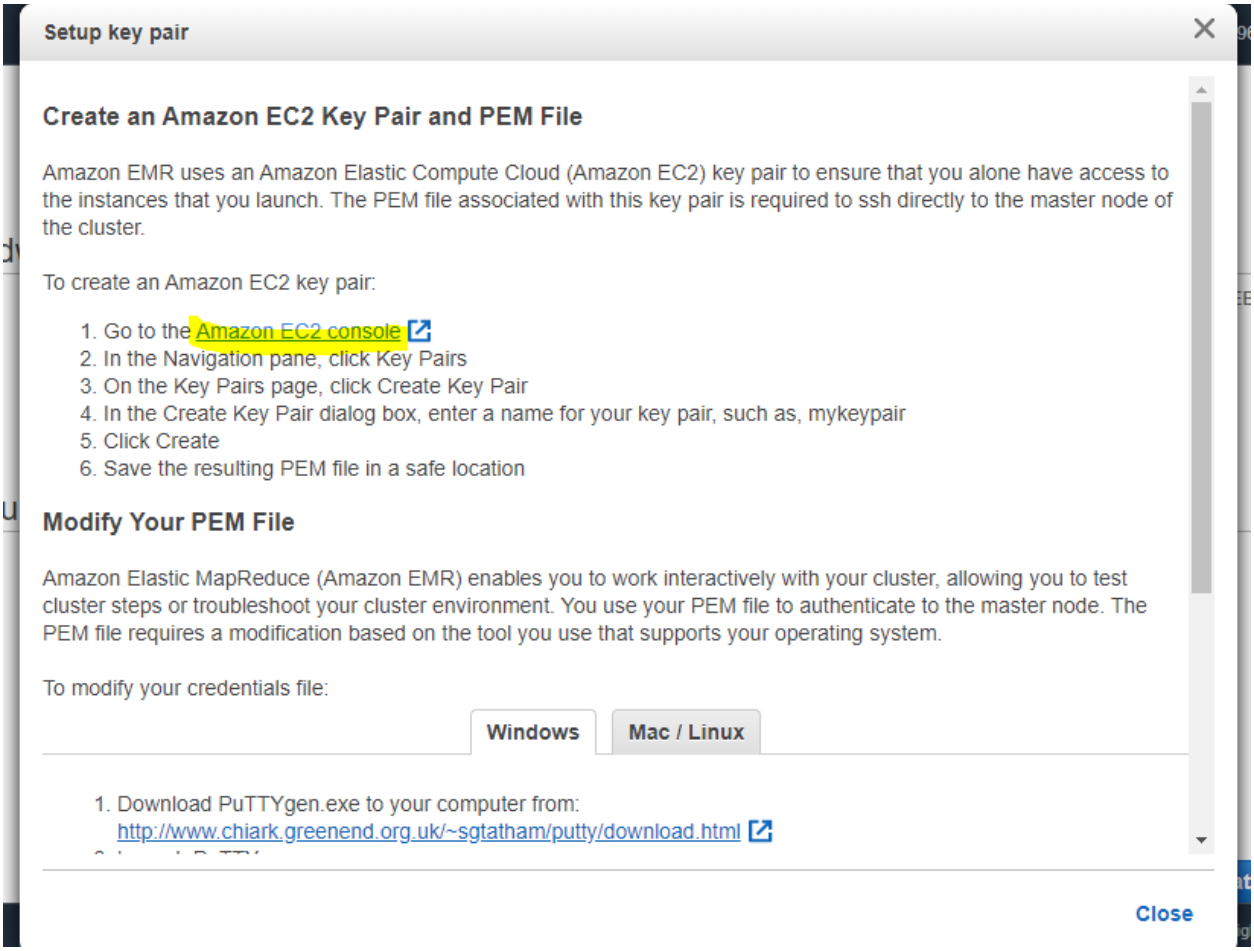
System Configurations

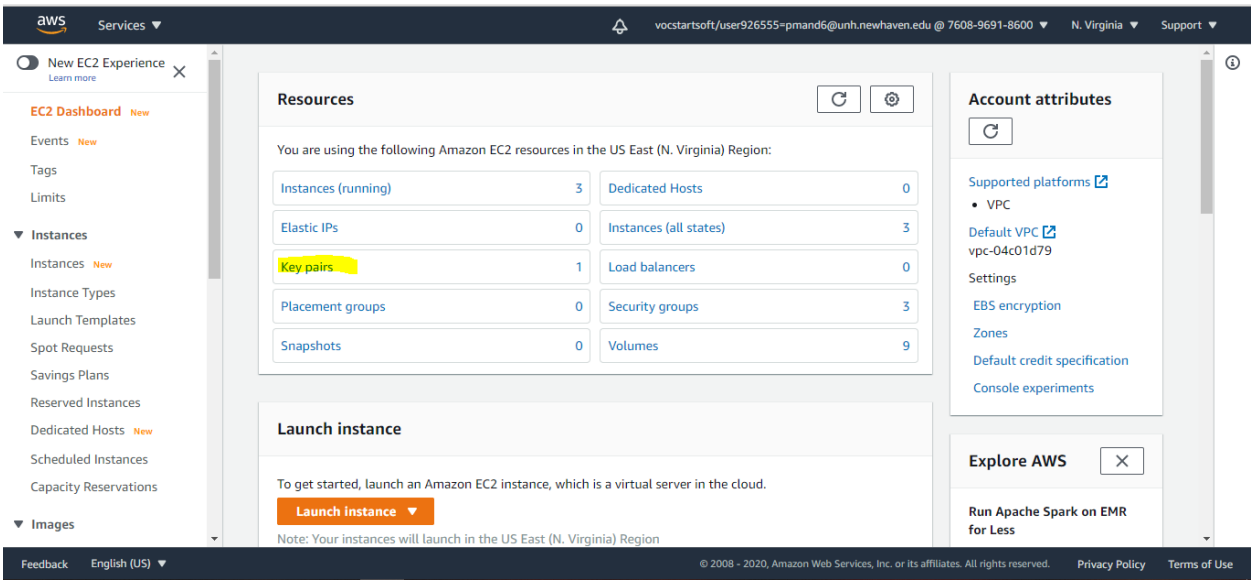


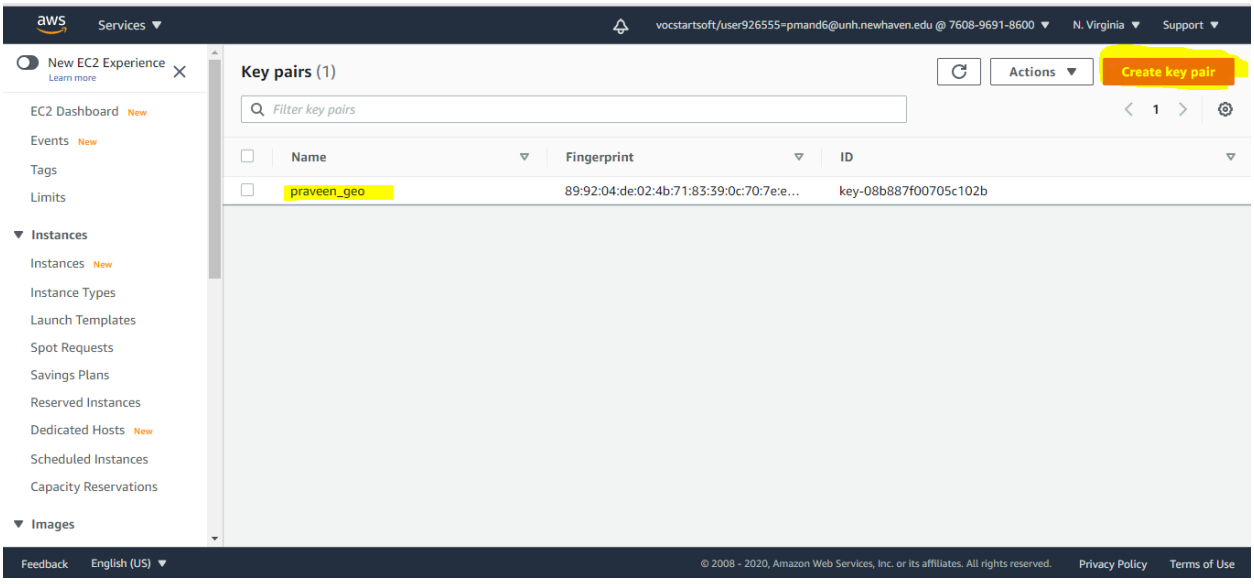

aws
Services
vocstartsoft/user926555=pmand6@unh.newhaven.edu @ 7608-9691-8600
N. Virginia
Support
Amazon EMR
Clusters
Notebooks
Git repositories
Security configurations
Block public access
VPC subnets
Events
Help
What's new
Save up to 90% when running your EMR clusters with EC2 Spot Instances. View tutorial
Create cluster
View details
Clone
Terminate
Filter: All clusters
Filter clusters ...
4 clusters (all loaded)
Name
ID
Status
Creation time (UTC-5)
Elapsed time
Normalized instance hou
My cluster
j-SN3ZQQ4XHH40
Waiting
Cluster ready
2020-12-07 02:06 (UTC-5)
1 day, 14 hours
912
My cluster
j-S3DDR4PB2DGK
Terminated with errors
Validation error
2020-12-07 02:05 (UTC-5)
24 seconds
0
My cluster
j-25V68DEJCN36I
Terminated
User request
2020-12-05 07:52 (UTC-5)
5 hours, 59 minutes
144
My cluster
j-2KJVYNC2KPXY4
Terminated
User request
2020-12-05 07:36 (UTC-5)
17 minutes
24
Feedback
English (US)
© 2008 - 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use

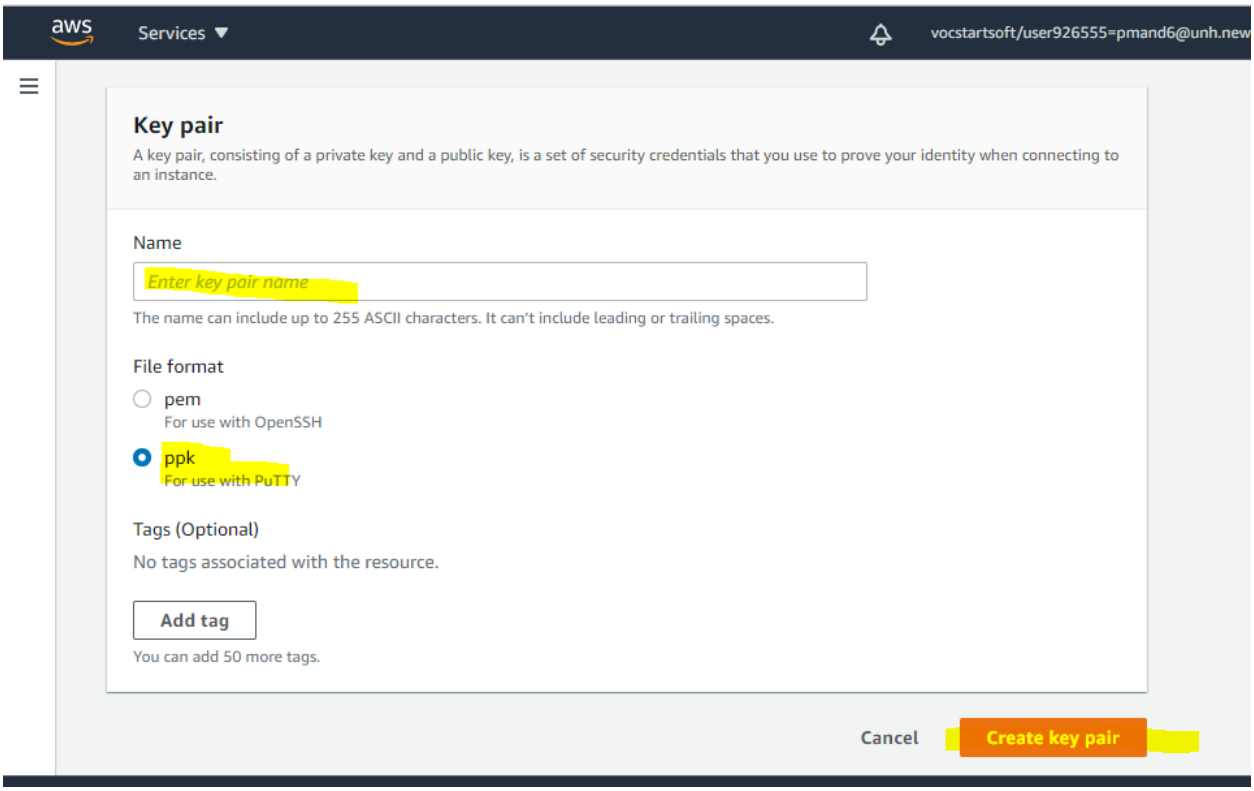

aws
Services
vocstartsoft/user926555=pmand6@unh.newhaven.edu @ 7608-9691-8600
N. Virginia
Support
Create Cluster - Quick Options
Go to advanced options
General Configuration
Cluster name
My cluster
Logging
S3 folder
s3://aws-logs-760896918600-us-east-1/elasticmapred
Launch mode
Cluster
Step execution
Software configuration
Release
emr-5.32.0
Applications
Core Hadoop: Hadoop 2.10.1, Hive 2.3.7, Hue 4.8.0, Mahout 0.13.0, Pig 0.17.0, and Tez 0.9.2
HBase: HBase 1.4.13, Hadoop 2.10.1, Hive 2.3.7, Hue 4.8.0, Phoenix 4.14.3, and ZooKeeper 3.4.14
Presto: Presto 0.240.1 with Hadoop 2.10.1 HDFS and Hive 2.3.7 Metastore
Spark: Spark 2.4.7 on Hadoop 2.10.1 YARN and Zeppelin 0.8.2
Use AWS Glue Data Catalog for table metadata
Feedback
English (US)
© 2008 - 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use

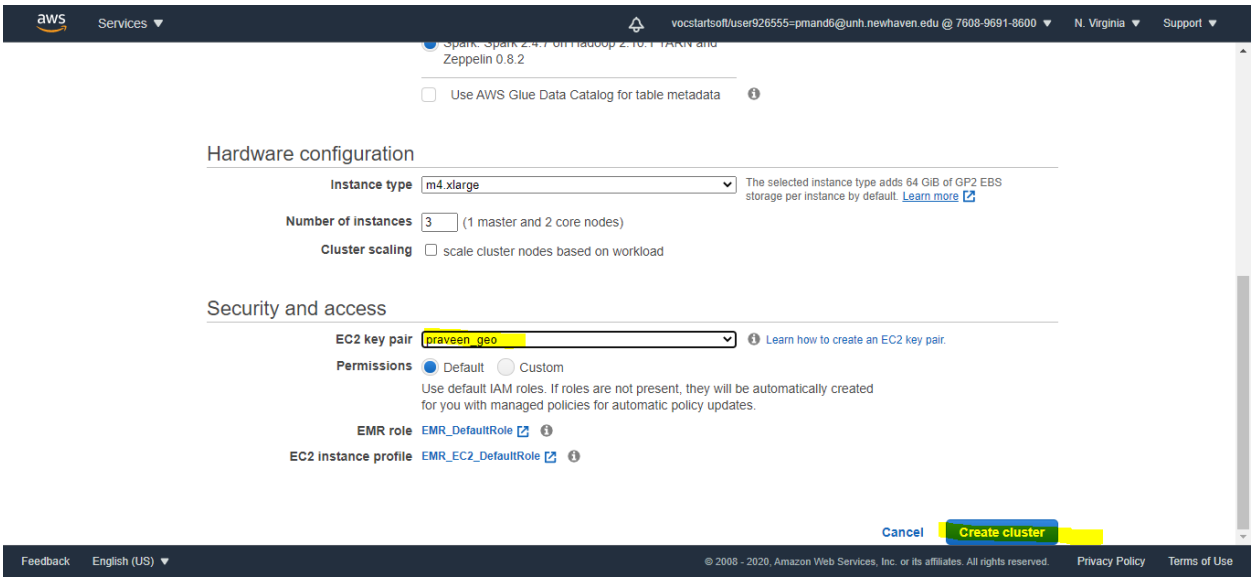

aws
Services
vocstartsoft/user926555=pmand6@unh.newhaven.edu @ 7608-9691-8600
N. Virginia
Support
Zeppelin 0.8.2
Use AWS Glue Data Catalog for table metadata
Hardware configuration
Instance type
m4.xlarge
The selected instance type adds 64 GiB of GP2 EBS storage per instance by default. Learn more
Number of instances
3
(1 master and 2 core nodes)
Cluster scaling
scale cluster nodes based on workload
Security and access
EC2 key pair
Choose an option
Learn how to create an EC2 key pair.
Permissions
Default
Custom
Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.
EMR role
EMR_DefaultRole
EC2 instance profile
EMR_EC2_DefaultRole
Cancel
Create cluster
Feedback
English (US)
© 2008 - 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use

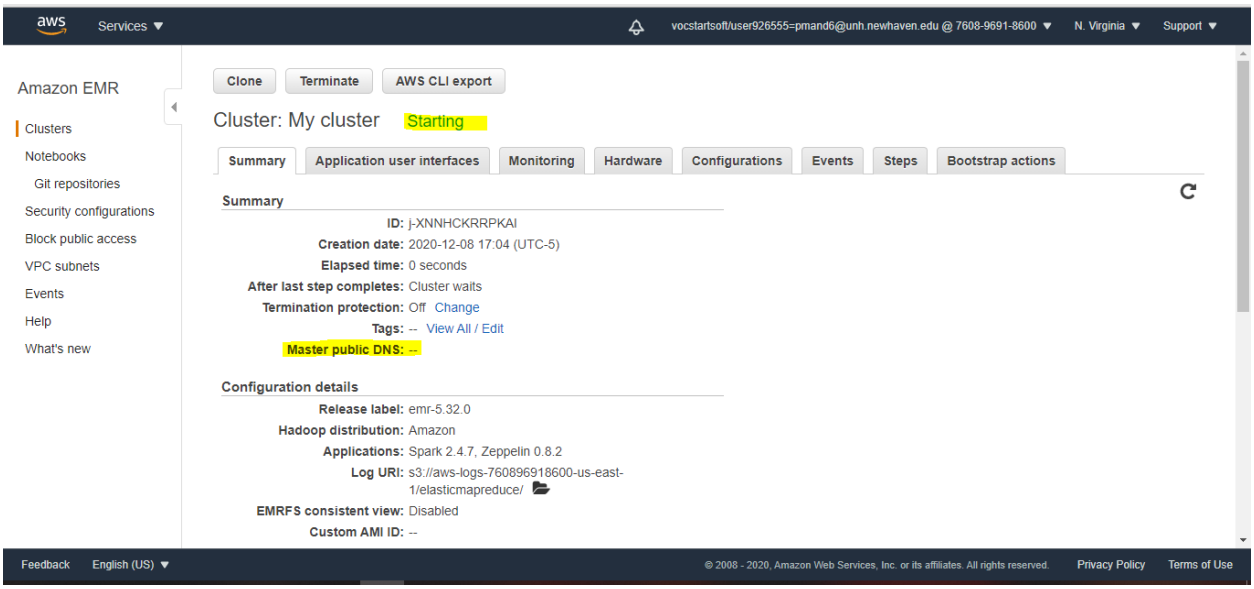

Setup key pair
Create an Amazon EC2 Key Pair and PEM File
Amazon EMR uses an Amazon Elastic Compute Cloud (Amazon EC2) key pair to ensure that you alone have access to the instances that you launch. The PEM file associated with this key pair is required to ssh directly to the master node of the cluster.
To create an Amazon EC2 key pair:
1. Go to the Amazon EC2 console
2. In the Navigation pane, click Key Pairs
3. On the Key Pairs page, click Create Key Pair
4. In the Create Key Pair dialog box, enter a name for your key pair, such as, mykeypair
5. Click Create
6. Save the resulting PEM file in a safe location
Modify Your PEM File
Amazon Elastic MapReduce (Amazon EMR) enables you to work interactively with your cluster, allowing you to test cluster steps or troubleshoot your cluster environment. You use your PEM file to authenticate to the master node. The PEM file requires a modification based on the tool you use that supports your operating system.
To modify your credentials file:
Windows
Mac / Linux
1. Download PuTTYgen.exe to your computer from:
http://www.chiark.greenend.org.uk/~sgtatham/putty/download.html
Close

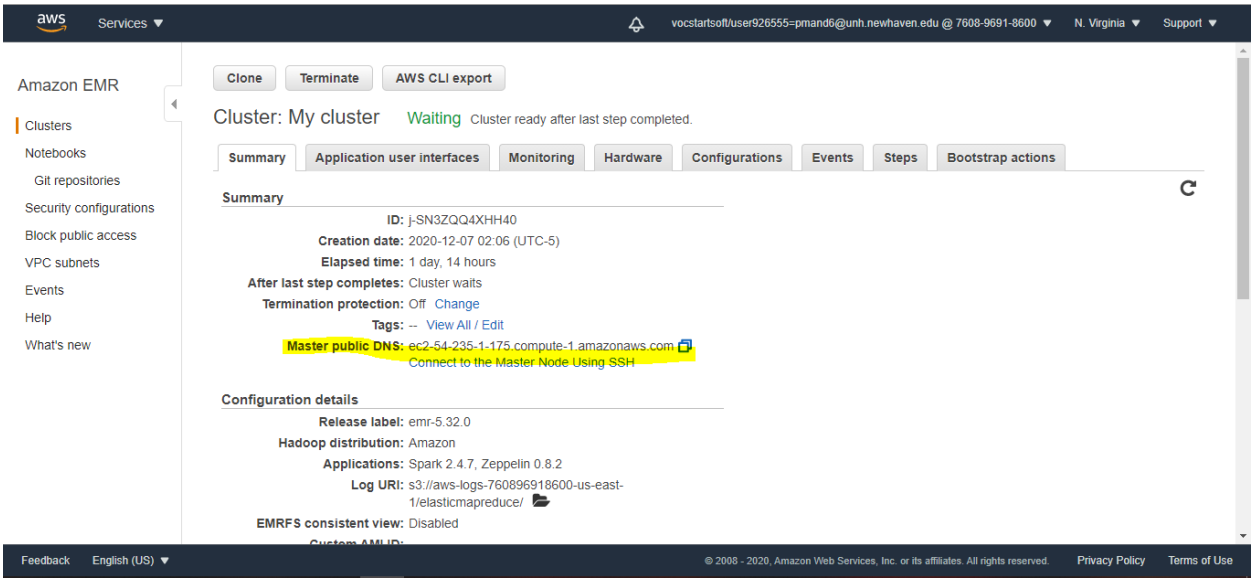

aws
Services
vocstartsoft/user926555=pmand6@unh.newhaven.edu @ 7608-9691-8600
N. Virginia
Support
New EC2 Experience
Learn more
EC2 Dashboard New
Events New
Tags
Limits
Instances
Instances New
Instance Types
Launch Templates
Spot Requests
Savings Plans
Reserved Instances
Dedicated Hosts New
Scheduled Instances
Capacity Reservations
Images
Resources
You are using the following Amazon EC2 resources in the US East (N. Virginia) Region:
Instances (running) 3
Dedicated Hosts 0
Elastic IPs 0
Instances (all states) 3
Key pairs 1
Load balancers 0
Placement groups 0
Security groups 3
Snapshots 0
Volumes 9
Launch instance
To get started, launch an Amazon EC2 instance, which is a virtual server in the cloud.
Launch instance
Note: Your instances will launch in the US East (N. Virginia) Region
Account attributes
Supported platforms
VPC
Default VPC
vpc-04c01d79
Settings
EBS encryption
Zones
Default credit specification
Console experiments
Explore AWS
Run Apache Spark on EMR for Less
Feedback
English (US)
© 2008 - 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use

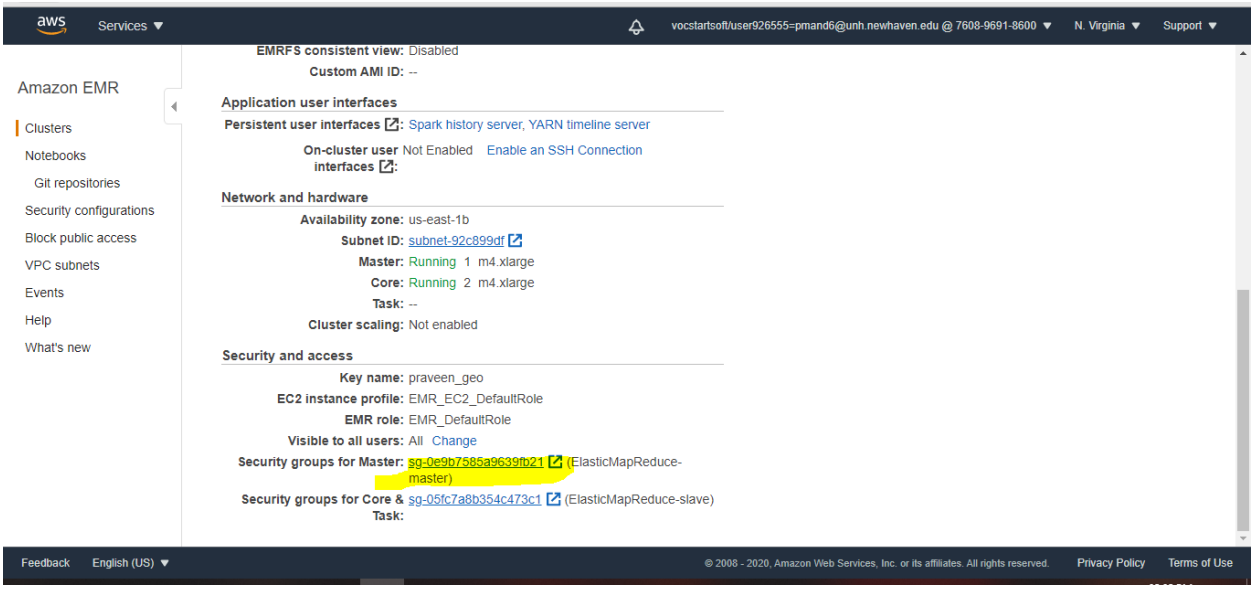

aws
Services
vocstartsoft/user926555=pmand6@unh.newhaven.edu @ 7608-9691-8600
N. Virginia
Support
New EC2 Experience
Learn more
EC2 Dashboard New
Events New
Tags
Limits
Instances
Instances New
Instance Types
Launch Templates
Spot Requests
Savings Plans
Reserved Instances
Dedicated Hosts New
Scheduled Instances
Capacity Reservations
Images
Key pairs (1)
Actions
Create key pair
Filter key pairs
1
Name
Fingerprint
ID
praveen_geo
89:92:04:de:02:4b:71:83:39:0c:70:7e:e...
key-08b887f00705c102b
Feedback
English (US)
© 2008 - 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use

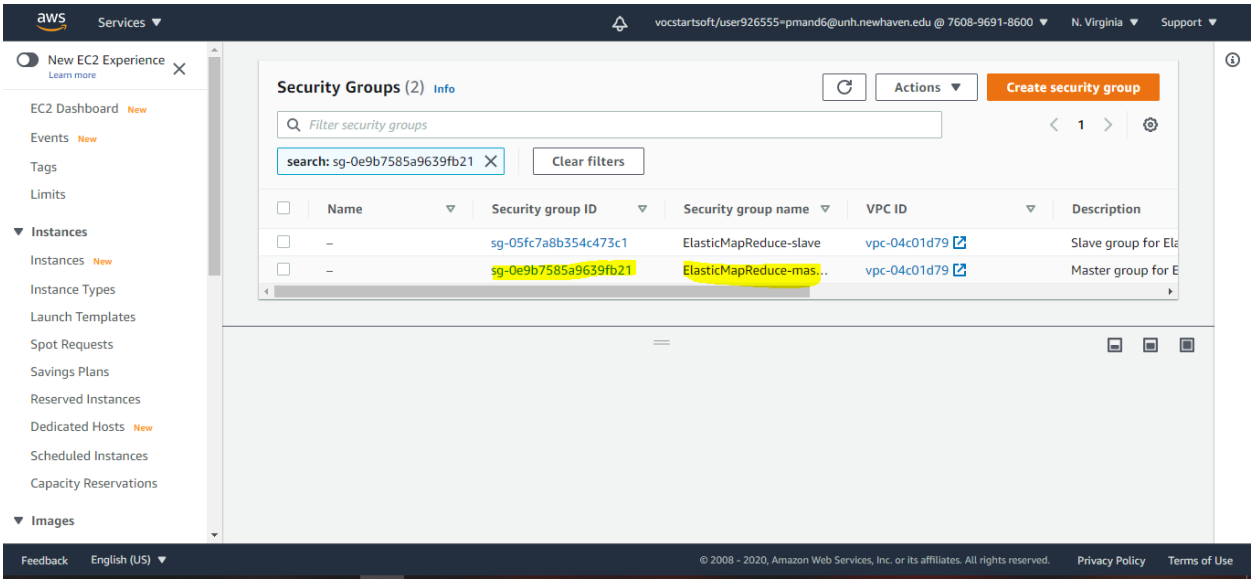

aws
Services
vocstartsoft/user926555=pmand6@unh.new
Key pair
A key pair, consisting of a private key and a public key, is a set of security credentials that you use to prove your identity when connecting to an instance.
Name
Enter key pair name
The name can include up to 255 ASCII characters. It can't include leading or trailing spaces.
File format
pem
For use with OpenSSH
ppk
For use with PuTTY
Tags (Optional)
No tags associated with the resource.
Add tag
You can add 50 more tags.
Cancel
Create key pair

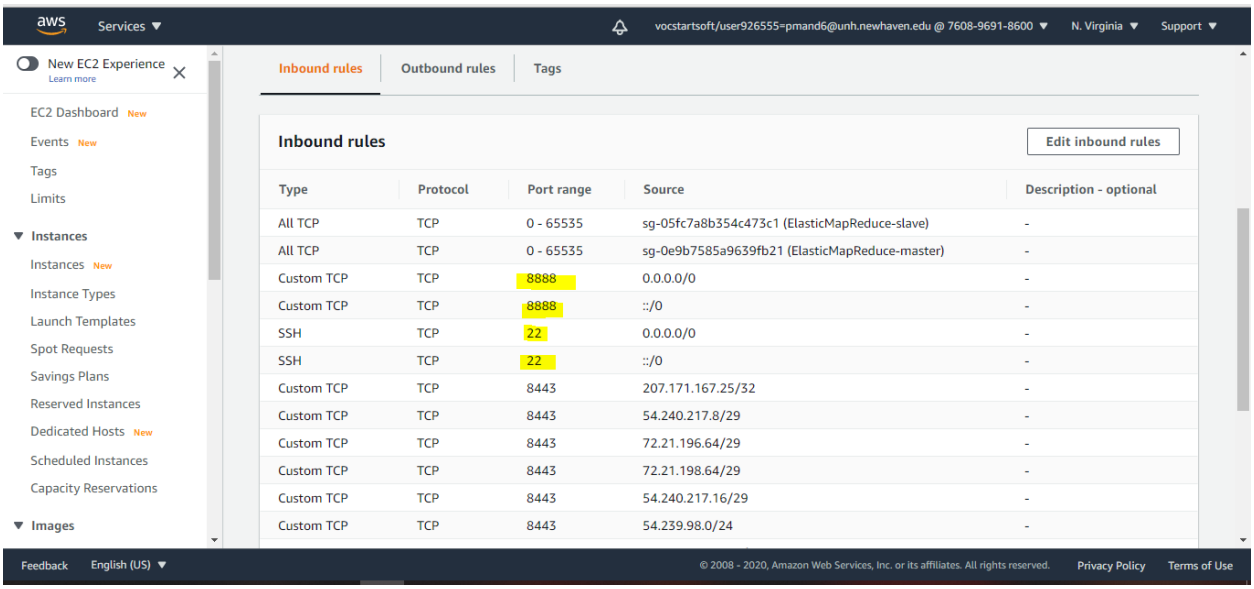

aws
Services
vocstartsoft/user926555=pmand6@unh.newhaven.edu @ 7608-9691-8600
N. Virginia
Support
Zeppelin 0.8.2
Use AWS Glue Data Catalog for table metadata
Hardware configuration
Instance type
m4.xlarge
The selected instance type adds 64 GiB of GP2 EBS storage per instance by default. Learn more
Number of instances
3
(1 master and 2 core nodes)
Cluster scaling
scale cluster nodes based on workload
Security and access
EC2 key pair
praveen_geo
Learn how to create an EC2 key pair.
Permissions
Default
Custom
Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.
EMR role
EMR_DefaultRole
EC2 instance profile
EMR_EC2_DefaultRole
Cancel
Create cluster
Feedback
English (US)
© 2008 - 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use

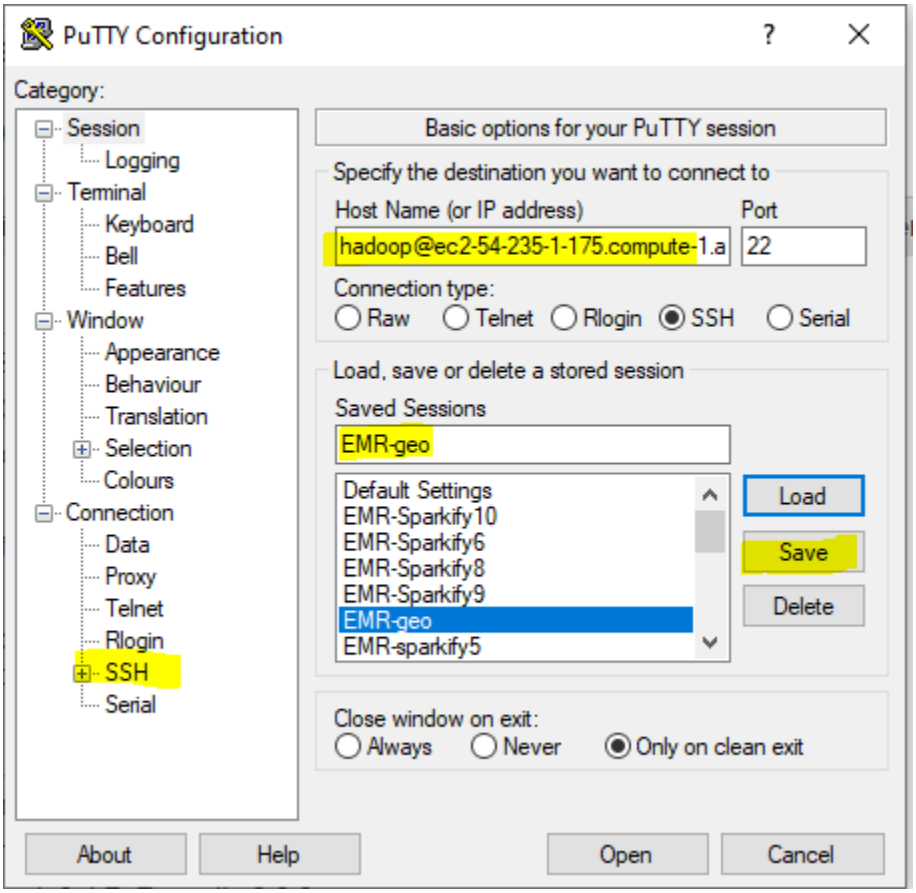

aws
Services
vocstartsoft/user926555=pmand6@unh.newhaven.edu @ 7608-9691-8600
N. Virginia
Support
Amazon EMR
Clusters
Notebooks
Git repositories
Security configurations
Block public access
VPC subnets
Events
Help
What's new
Clone
Terminate
AWS CLI export
Cluster: My cluster
Starting
Summary
Application user interfaces
Monitoring
Hardware
Configurations
Events
Steps
Bootstrap actions
Summary
ID: j-XNNHCKRRPKAI
Creation date: 2020-12-08 17:04 (UTC-5)
Elapsed time: 0 seconds
After last step completes: Cluster waits
Termination protection: Off Change
Tags: -- View All / Edit
Master public DNS: --
Configuration details
Release label: emr-5.32.0
Hadoop distribution: Amazon
Applications: Spark 2.4.7, Zeppelin 0.8.2
Log URI: s3://aws-logs-760896918600-us-east-1/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --
Feedback
English (US)
© 2008 - 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use

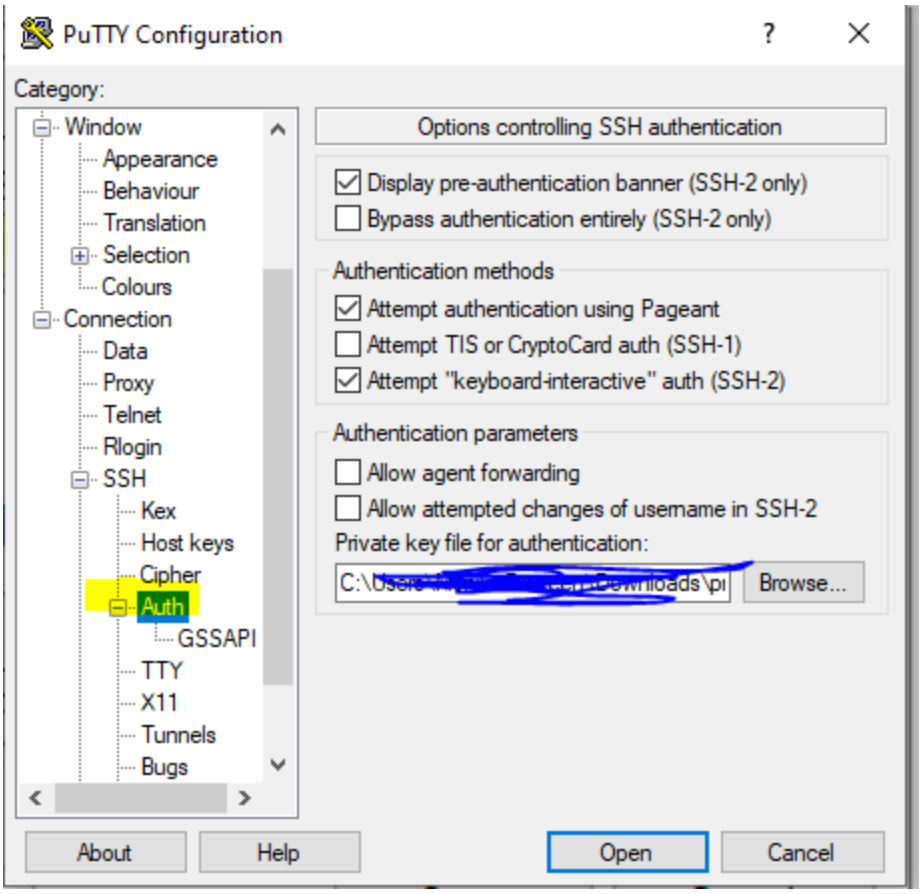

aws
Services
vocstartsoft/user926555=pmand6@unh.newhaven.edu @ 7608-9691-8600
N. Virginia
Support
Amazon EMR
Clusters
Notebooks
Git repositories
Security configurations
Block public access
VPC subnets
Events
Help
What's new
Clone
Terminate
AWS CLI export
Cluster: My cluster
Waiting
Cluster ready after last step completed.
Summary
Application user interfaces
Monitoring
Hardware
Configurations
Events
Steps
Bootstrap actions
Summary
ID: j-SN3ZQQ4XHH40
Creation date: 2020-12-07 02:06 (UTC-5)
Elapsed time: 1 day, 14 hours
After last step completes: Cluster waits
Termination protection: Off Change
Tags: -- View All / Edit
Master public DNS: ec2-54-235-1-175.compute-1.amazonaws.com
Connect to the Master Node Using SSH
Configuration details
Release label: emr-5.32.0
Hadoop distribution: Amazon
Applications: Spark 2.4.7, Zeppelin 0.8.2
Log URI: s3://aws-logs-760896918600-us-east-1/elasticmapreduce/
EMRFS consistent view: Disabled
Feedback
English (US)
© 2008 - 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use


aws
Services
vocstartsoft/user926555=pmand6@unh.newhaven.edu @ 7608-9691-8600
N. Virginia
Support
Amazon EMR
Clusters
Notebooks
Git repositories
Security configurations
Block public access
VPC subnets
Events
Help
What's new
EMRFS consistent view: Disabled
Custom AMI ID: --
Application user interfaces
Persistent user interfaces: Spark history server, YARN timeline server
On-cluster user interfaces: Not Enabled Enable an SSH Connection
Network and hardware
Availability zone: us-east-1b
Subnet ID: subnet-92c899df
Master: Running 1 m4.xlarge
Core: Running 2 m4.xlarge
Task: --
Cluster scaling: Not enabled
Security and access
Key name: praveen_geo
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Visible to all users: All Change
Security groups for Master: sg-0e9b7585a9639fb21 (ElasticMapReduce-master)
Security groups for Core & Task: sg-05fc7a8b354c473c1 (ElasticMapReduce-slave)
Feedback
English (US)
© 2008 - 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use


aws
Services
vocstartsoft/user926555=pmand6@unh.newhaven.edu @ 7608-9691-8600
N. Virginia
Support
New EC2 Experience
Learn more
EC2 Dashboard New
Events New
Tags
Limits
Instances
Instances New
Instance Types
Launch Templates
Spot Requests
Savings Plans
Reserved Instances
Dedicated Hosts New
Scheduled Instances
Capacity Reservations
Images
Security Groups (2) Info
Actions
Create security group
Filter security groups
search: sg-0e9b7585a9639fb21
Clear filters
Name
Security group ID
Security group name
VPC ID
Description
sg-05fc7a8b354c473c1
ElasticMapReduce-slave
vpc-04c01d79
sg-0e9b7585a9639fb21
ElasticMapReduce-mas...
vpc-04c01d79
Feedback
English (US)
© 2008 - 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use


aws
Services
vocstartsoft/user926555=pmand6@unh.newhaven.edu @ 7608-9691-8600
N. Virginia
Support
New EC2 Experience
Learn more
EC2 Dashboard New
Events New
Tags
Limits
Instances
Instances New
Instance Types
Launch Templates
Spot Requests
Savings Plans
Reserved Instances
Dedicated Hosts New
Scheduled Instances
Capacity Reservations
Images
Inbound rules
Outbound rules
Tags
Inbound rules
Edit inbound rules
Type Protocol Port range Source Description - optional
All TCP TCP 0 - 65535 sg-05fc7a8b354c473c1 (ElasticMapReduce-slave) -
All TCP TCP 0 - 65535 sg-0e9b7585a9639fb21 (ElasticMapReduce-master) -
Custom TCP TCP 8888 0.0.0.0/0 -
Custom TCP TCP 8888 ::/0 -
SSH TCP 22 0.0.0.0/0 -
SSH TCP 22 ::/0 -
Custom TCP TCP 8443 207.171.167.25/32 -
Custom TCP TCP 8443 54.240.217.8/29 -
Custom TCP TCP 8443 72.21.196.64/29 -
Custom TCP TCP 8443 72.21.198.64/29 -
Custom TCP TCP 8443 54.240.217.16/29 -
Custom TCP TCP 8443 54.239.98.0/24 -
Feedback
English (US)
© 2008 - 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use


PuTTY Configuration
Category:
Session
Logging
Terminal
Keyboard
Bell
Features
Window
Appearance
Behaviour
Translation
Selection
Colours
Connection
Data
Proxy
Telnet
Rlogin
SSH
Serial
Basic options for your PuTTY session
Specify the destination you want to connect to
Host Name (or IP address)
Port
hadoop@ec2-54-235-1-175.compute-1.a
22
Connection type:
Raw
Telnet
Rlogin
SSH
Serial
Load, save or delete a stored session
Saved Sessions
EMR-geo
Default Settings
EMR-Sparkify10
EMR-Sparkify6
EMR-Sparkify8
EMR-Sparkify9
EMR-geo
EMR-sparkify5
Load
Save
Delete
Close window on exit:
Always
Never
Only on clean exit
About
Help
Open
Cancel


PuTTY Configuration
Category:
Window
Appearance
Behaviour
Translation
Selection
Colours
Connection
Data
Proxy
Telnet
Rlogin
SSH
Kex
Host keys
Cipher
Auth
GSSAPI
TTY
X11
Tunnels
Bugs
Options controlling SSH authentication
Display pre-authentication banner (SSH-2 only)
Bypass authentication entirely (SSH-2 only)
Authentication methods
Attempt authentication using Pageant
Attempt TIS or CryptoCard auth (SSH-1)
Attempt "keyboard-interactive" auth (SSH-2)
Authentication parameters
Allow agent forwarding
Allow attempted changes of username in SSH-2
Private key file for authentication:
Browse...
About
Help
Open
Cancel

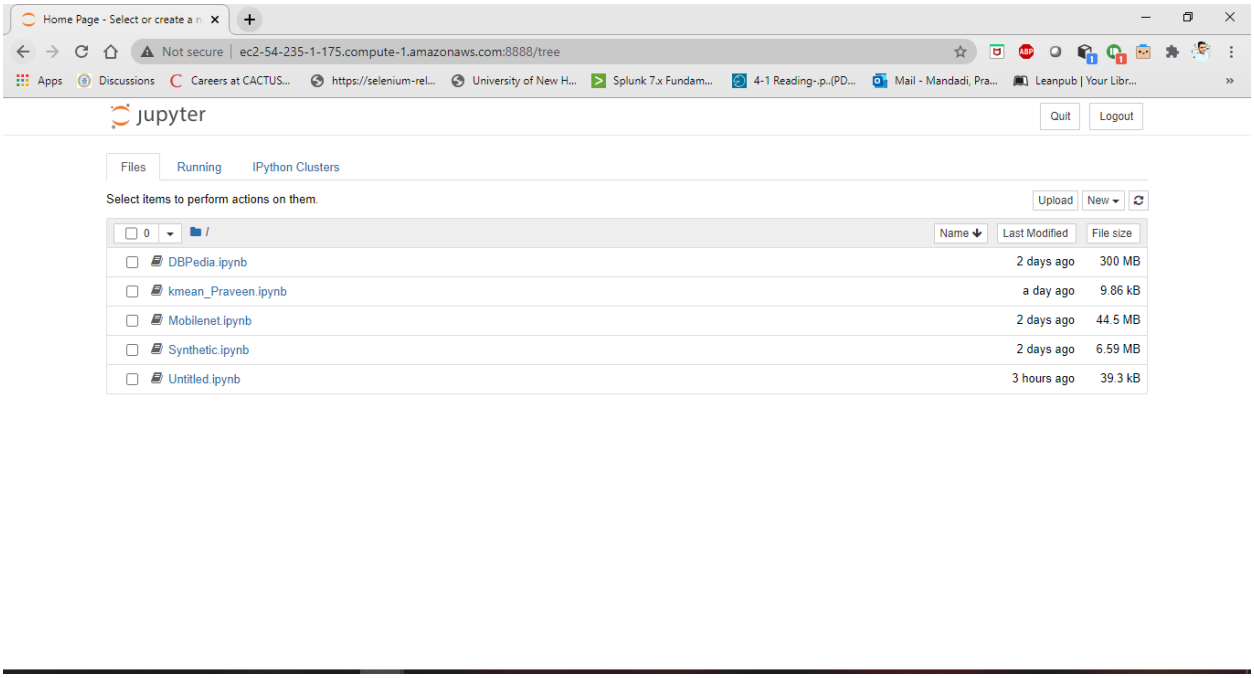
hadoop@ip-172-31-18-226:~

```
Using username "hadoop".
Authenticating with public key "praveen_geo"
Last login: Tue Dec  8 21:52:22 2020

       __|  __|_  )
       _|  (     /   Amazon Linux 2 AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-2/

EEEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-18-226 ~]$
```

hadoop@ip-172-31-18-226:~

```
# Source global definitions
if [ -f /etc/bashrc ]; then
  . /etc/bashrc
fi

_set_aws_region() {
  local curl_opts="--retry 5 -f --silent --connect-timeout 2"
  local token=$(curl ${curl_opts} -X PUT "http://169.254.169.254/latest/api/token" -H "X-aws-ec2-metadata-token-ttl-seconds: 15")
  export AWS_DEFAULT_REGION=$(curl ${curl_opts} -H "X-aws-ec2-metadata-token: ${token}" http://169.254.169.254/latest/dynamic/instance-identity/document | grep region |
 awk -F\" '{print $4}')

  unset -f $FUNCNAME
}

# set the default region for the AWS CLI
_set_aws_region
export JAVA_HOME=/etc/alternatives/jre
export HOME=/home/hadoop

export PYSPARK_DRIVER_PYTHON=/usr/local/bin/jupyter
export PYSPARK_DRIVER_PYTHON_OPTS="notebook --no-browser  --ip=0.0.0.0 --port=8888"
~
".bashrc" 20L, 763C                                                    20,83         All
```


hadoop@ip-172-31-18-226:~
Using username "hadoop".
Authenticating with public key "praveen_geo"
Last login: Tue Dec 8 21:52:22 2020
Amazon Linux 2 AMI
https://aws.amazon.com/amazon-linux-2/
[hadoop@ip-172-31-18-226 ~]$ sudo vim .bashrc
[hadoop@ip-172-31-18-226 ~]$
[hadoop@ip-172-31-18-226 ~]$
[hadoop@ip-172-31-18-226 ~]$ pyspark
[I 22:11:14.715 NotebookApp] Loading IPython parallel extension
[I 22:11:14.716 NotebookApp] Serving notebooks from local directory: /home/hadoop
[I 22:11:14.716 NotebookApp] Jupyter Notebook 6.1.5 is running at:
[I 22:11:14.716 NotebookApp] http://ip-172-31-18-226:8888/?token=09d97d3955383e90dfa3d7917a3442ebb7ed025f751887c9
[I 22:11:14.716 NotebookApp] or http://127.0.0.1:8888/?token=09d97d3955383e90dfa3d7917a3442ebb7ed025f751887c9
[I 22:11:14.716 NotebookApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
[C 22:11:14.720 NotebookApp]
To access the notebook, open this file in a browser:
file:///home/hadoop/.local/share/jupyter/runtime/nbserver-20088-open.html
Or copy and paste one of these URLs:
http://ip-172-31-18-226:8888/?token=09d97d3955383e90dfa3d7917a3442ebb7ed025f751887c9
or http://127.0.0.1:8888/?token=09d97d3955383e90dfa3d7917a3442ebb7ed025f751887c9


Home Page - Select or create a n
Not secure | ec2-54-235-1-175.compute-1.amazonaws.com:8888/tree
Apps
Discussions
Careers at CACTUS...
https://selenium-rel...
University of New H...
Splunk 7.x Fundam...
4-1 Reading-.p..(PD...
Mail - Mandadi, Pra...
Leanpub | Your Libr...
jupyter
Quit
Logout
Files
Running
IPython Clusters
Select items to perform actions on them.
Upload
New
Name
Last Modified
File size
DBPedia.ipynb 2 days ago 300 MB
kmean_Praveen.ipynb a day ago 9.86 kB
Mobilenet.ipynb 2 days ago 44.5 MB
Synthetic.ipynb 2 days ago 6.59 MB
Untitled.ipynb 3 hours ago 39.3 kB

Data Pre-Processing

We went through pre-processing steps before implementing the real algorithm to translate the data for later processing into a standardized format. The pre-processing process for device status data is listed below.

A. Load the dataset.
B. Determine which delimiter to use.
C. Filter out any records which do not parse correctly; each record should have exactly 14 values.
D. Extract the date, model, device ID, and latitude and longitude.
    I. date: 1st field
    II. model: 2nd field
    III. device ID: 3rd field
    IV. latitude: 13th field
    V. longitude: 14th field
E. Store latitude and longitude as the first two fields
F. Filter out locations that have a latitude and longitude of 0.
G. Split the model field that contains the device manufacturer and model name by spaces.
H. Save the extracted data as comma separated values file in the s3://datageo1/results/devicedata.csv directory on AWS S3.
I. Confirm the data in the file(s) was saved correctly.

Device Status dataset

In [16]:
```
#Printing the data frame after droping zeros
devicedata
```

Out[16]:

	latitude	longitude	date	model	device ID
0	33.6894754264	-117.543308253	2014-03-15:10:10:20	F41L	8cc3b47e-bd01-4482-b500-28f2342679af
1	39.3635186767	-119.400334708	2014-03-15:10:10:20	F41L	707daba1-5640-4d60-a6d9-1d6fa0645be0
2	33.1913581092	-116.448242643	2014-03-15:10:10:20	Novelty Note 1	db66fe81-aa55-43b4-9418-fc6e7a00f891
3	33.8343543748	-117.330000857	2014-03-15:10:10:20	F41L	ffa18088-69a0-433e-84b8-006b2b9cc1d0
4	37.3803954321	-121.840756755	2014-03-15:10:10:20	F33L	66d678e6-9c87-48d2-a415-8d5035e54a23
...	...	...	...	...	...
65640	39.4463417571	-114.736213453	2014-03-15:10:49:30	F22L	40e61459-5448-4dc9-bb89-42e73a4e19cf
65641	38.4282665514	-121.25933863	2014-03-15:10:49:30	S2	b13ece99-62ab-4c9f-a366-6a06bd5e877f
65642	33.7778202246	-108.575470704	2014-03-15:10:49:30	F41L	32af1a0b-ca7f-4906-9772-9eb9435e7e4c
65643	38.2596913494	-122.295712621	2014-03-15:10:49:30	S1	a48a5559-d916-481b-84a9-5dce6272cce1
65644	34.2415255221	-118.23526739	2014-03-15:10:49:30	2	d86fbaa6-b71b-435f-a0bf-5304a202a70b

65645 rows × 5 columns

Synthetic Dataset

```
In [26]: #Displaying Pandas Data Frame
         pandas_df.head()
```

Out[26]:

	Latitude	Longitude	LocationID
0	37.77253945	-77.49954987	1
1	42.09013298	-87.68915558	2
2	39.56341754	-75.58753204	3
3	39.45302347	-87.69374084	4
4	38.9537989	-77.01656342	5

DBPedia Dataset

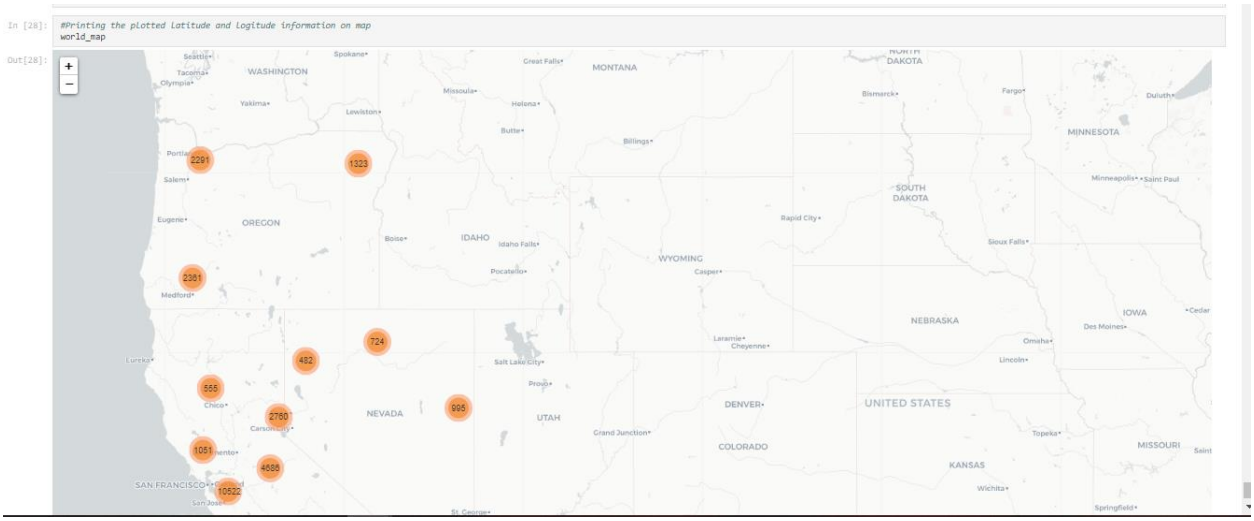
```
In [24]: #Showing pandas data frame
         pandas_df.head()
```

Out[24]:

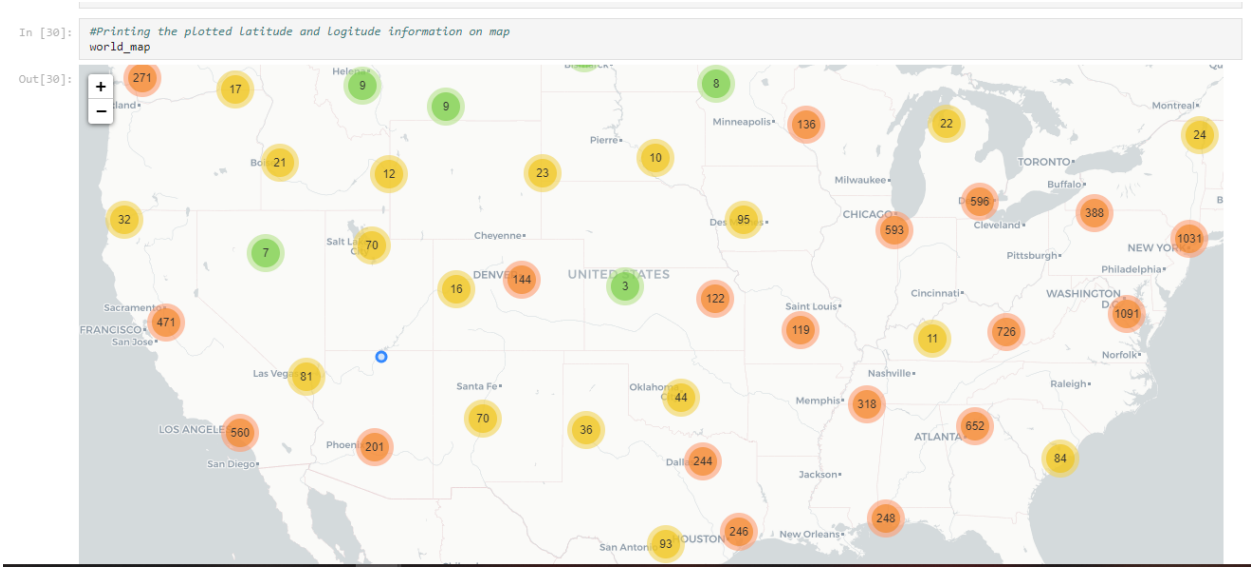
	lat	long	name_of_page
0	36.7	3.216666666666667	<http://dbpedia.org/resource/Algeria>
1	42.5	1.5166666666666666	<http://dbpedia.org/resource/Andorra>
2	12.516666666666667	-70.03333333333333	<http://dbpedia.org/resource/Aruba>
3	-8.833333333333334	13.333333333333334	<http://dbpedia.org/resource/Angola>
4	41.333333333333336	19.8	<http://dbpedia.org/resource/Albania>

Visualization

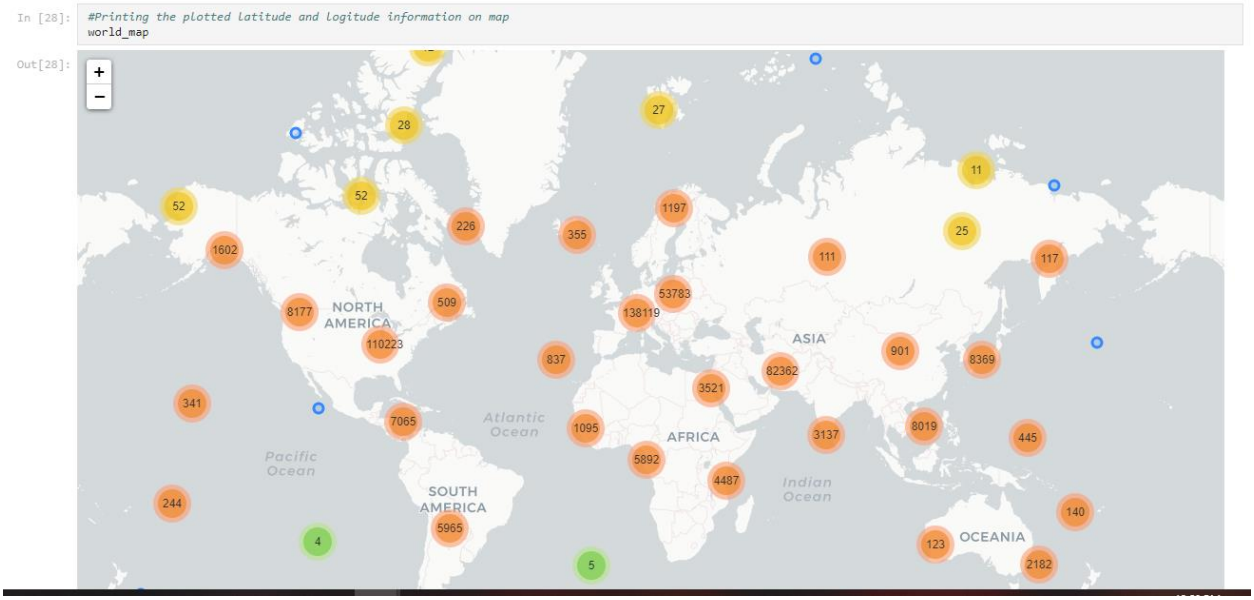
Mobile Net data



Synthetic Location Data



DBPedia Location Data



Clustering Approach

In the dataset, the initial centroids are a k-sized random sample of all items. The algorithm assigns each point to its nearest centroid for each iteration, then calculates the new centroids by taking the average of all points in the cluster of that centroid. Using either Euclidean distance or Great Circle distance, the distance between points and centroids is determined - the user sets this parameter. The key difference between the two measurements is that the former measures a straight-line distance between the points in 3D space, while the latter measures the distance around the spherical surface of the Planet.

Though a “perfect” algorithm would iterate until the change in centroid locations converges to 0, this algorithm continues iterating until the sum of all changes in centroid locations converges to α=0.1 km. That is, the algorithm calculates the distance between the current position of each centroid and the former location (using the user-specified distance measure) for each iteration and proceeds to iterate until the sum of these distances is less than 0.1 km for all centroids. This requires larger values of k to converge more precisely than smaller values. Because data this algorithm runs on covers at least an entire continent, we determined that the alpha of 0.1 km or 100 m was a small enough value for this purpose.