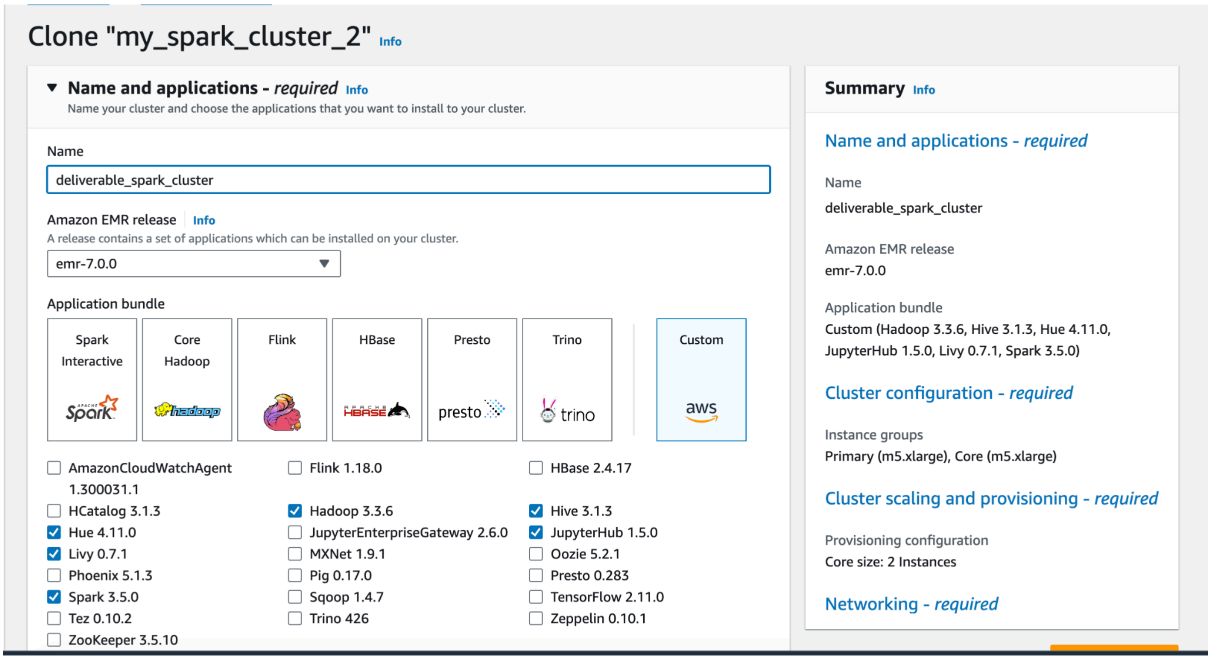
Big Data Wrangling With Google Books Ngrams

Author: Tilova Shahrin

1. On AWS Management Console, click on the search bar and type ‘EMR’. Click on EMR.
   a. Cloning cluster from spark lab.
   b. My_spark_cluster_2 is the cluster we will clone from, since it has all the features we need from spark lab.
   c.



Clone "my_spark_cluster_2" Info

Name and applications - *required* Info

Name your cluster and choose the applications that you want to install to your cluster.

Name

deliverable_spark_cluster

Amazon EMR release Info

A release contains a set of applications which can be installed on your cluster.

emr-7.0.0

Application bundle

□ AmazonCloudWatchAgent 1.300031.1
□ HCatalog 3.1.3
☑ Hue 4.11.0
☑ Livy 0.7.1
□ Phoenix 5.1.3
☑ Spark 3.5.0
□ Tez 0.10.2
□ ZooKeeper 3.5.10
□ Flink 1.18.0
☑ Hadoop 3.3.6
□ JupyterEnterpriseGateway 2.6.0
□ MXNet 1.9.1
□ Pig 0.17.0
□ Sqoop 1.4.7
□ Trino 426
□ HBase 2.4.17
☑ Hive 3.1.3
☑ JupyterHub 1.5.0
□ Oozie 5.2.1
□ Presto 0.283
□ TensorFlow 2.11.0
□ Zeppelin 0.10.1

Summary Info

Name and applications - *required*

Name
deliverable_spark_cluster

Amazon EMR release
emr-7.0.0

Application bundle
Custom (Hadoop 3.3.6, Hive 3.1.3, Hue 4.11.0, JupyterHub 1.5.0, Livy 0.7.1, Spark 3.5.0)

Cluster configuration - *required*

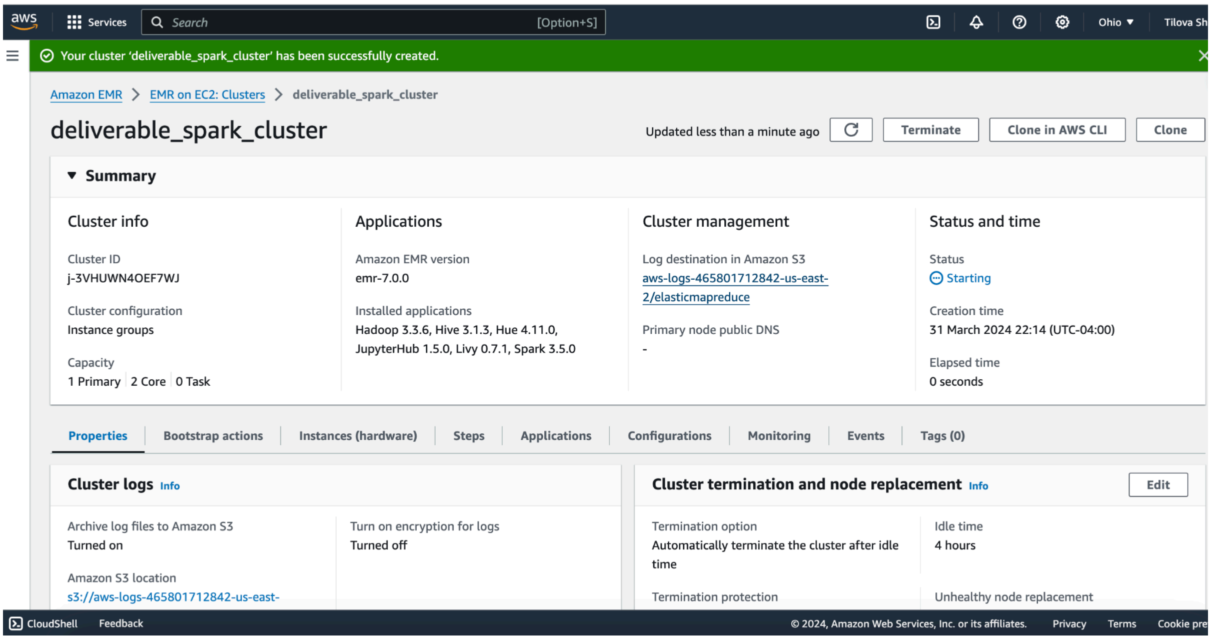
Instance groups
Primary (m5.xlarge), Core (m5.xlarge)

Cluster scaling and provisioning - *required*

Provisioning configuration
Core size: 2 Instances

Networking - *required*

   d. Click on ‘Clone Cluster’
   e.



Your cluster ‘deliverable_spark_cluster’ has been successfully created.

Amazon EMR > EMR on EC2: Clusters > deliverable_spark_cluster

deliverable_spark_cluster

Updated less than a minute ago

Summary

Cluster info

Cluster ID
j-3VHUWN4OEF7WJ

Cluster configuration
Instance groups

Capacity
1 Primary | 2 Core | 0 Task

Applications

Amazon EMR version
emr-7.0.0

Installed applications
Hadoop 3.3.6, Hive 3.1.3, Hue 4.11.0, JupyterHub 1.5.0, Livy 0.7.1, Spark 3.5.0

Cluster management

Log destination in Amazon S3
aws-logs-465801712842-us-east-2/elasticmapreduce

Primary node public DNS
-

Status and time

Status
Starting

Creation time
31 March 2024 22:14 (UTC-04:00)

Elapsed time
0 seconds

Cluster logs Info

Archive log files to Amazon S3
Turned on

Turn on encryption for logs
Turned off

Amazon S3 location
s3://aws-logs-465801712842-us-east-

Cluster termination and node replacement Info

Termination option
Automatically terminate the cluster after idle time

Idle time
4 hours

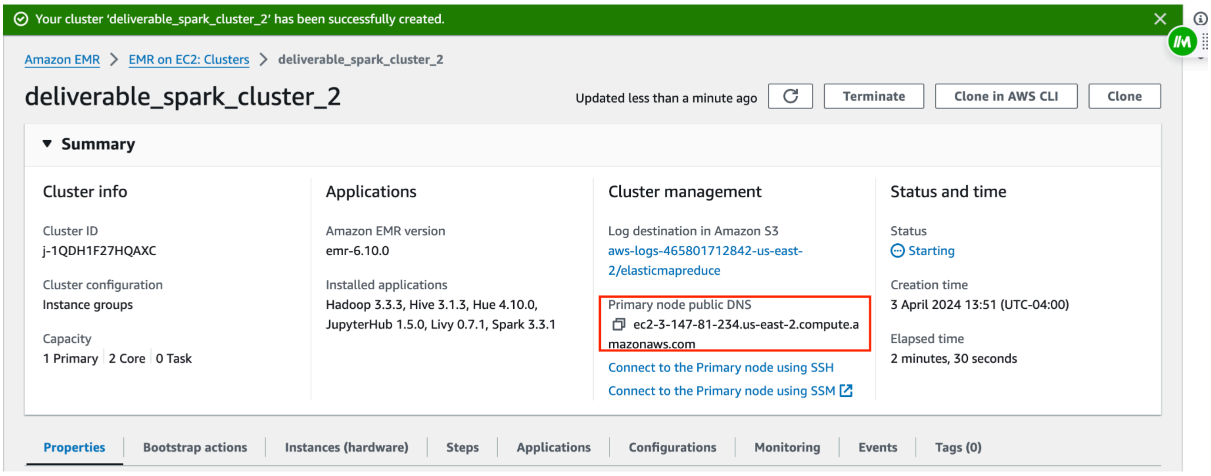
Termination protection

Unhealthy node replacement

2. Connect to the head node of the cluster using SSH.
   a. In your terminal, make sure you are in the same directory as your .pem file.
   b. Follow these instructions under ‘Connect to the primary node using SSH’. Paste the command onto the terminal. You can find your primary node public DNS in your cluster page, see the following command and image.

c. *ssh -i aws_cloud_tilova.pem -L 9995:localhost:9443 hadoop@ec2-18-118-27-195.us-east-2.compute.amazonaws.com*

d.



Your cluster 'deliverable_spark_cluster_2' has been successfully created.

Amazon EMR > EMR on EC2: Clusters > deliverable_spark_cluster_2

deliverable_spark_cluster_2

Updated less than a minute ago

▼ Summary

Cluster info

Cluster ID
j-1QDH1F27HQAXC

Cluster configuration
Instance groups

Capacity
1 Primary | 2 Core | 0 Task

Applications

Amazon EMR version
emr-6.10.0

Installed applications
Hadoop 3.3.3, Hive 3.1.3, Hue 4.10.0, JupyterHub 1.5.0, Livy 0.7.1, Spark 3.3.1

Cluster management

Log destination in Amazon S3
aws-logs-465801712842-us-east-2/elasticmapreduce

Primary node public DNS
ec2-3-147-81-234.us-east-2.compute.amazonaws.com

Connect to the Primary node using SSH
Connect to the Primary node using SSM

Status and time

Status
Starting

Creation time
3 April 2024 13:51 (UTC-04:00)

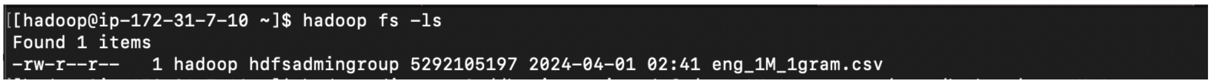
Elapsed time
2 minutes, 30 seconds

3. Copy the data folder from the S3 bucket *directly* into a directory on the Hadoop File System (HDFS) named `/user/hadoop/eng_1M_1gram`.

a. Add this command onto the terminal: *hadoop distcp s3://brainstation-dsft/eng_1M_1gram.csv /user/hadoop/eng_1M_1gram*.

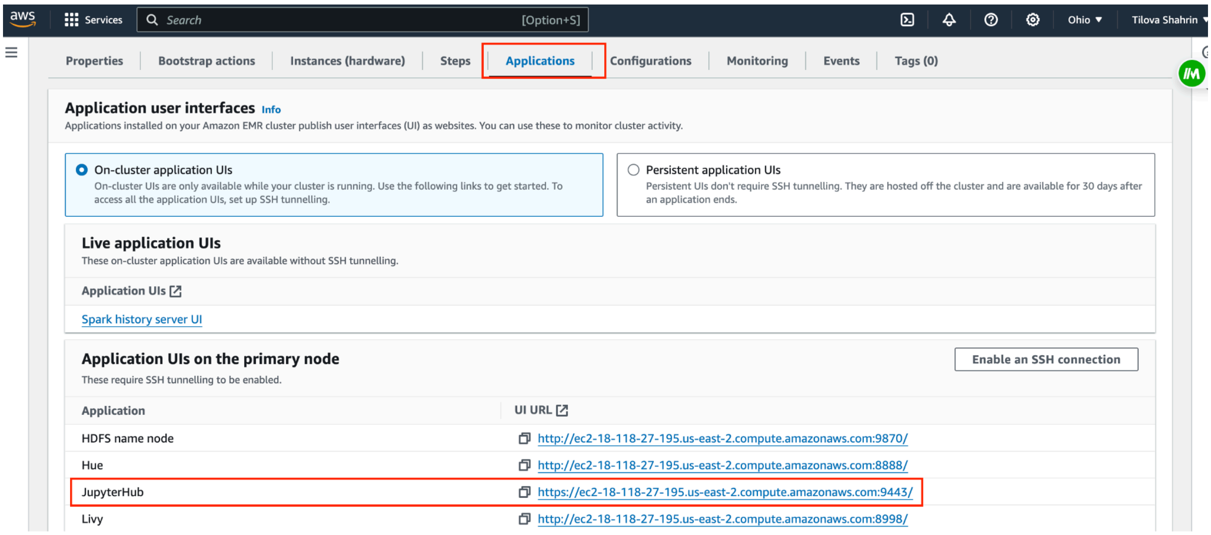
b. You can double check if your file exists in terminal using *hadoop fs -ls*

c.



```
[hadoop@ip-172-31-7-10 ~]$ hadoop fs -ls
Found 1 items
-rw-r--r--   1 hadoop hdfsadmingroup 5292105197 2024-04-01 02:41 eng_1M_1gram.csv
```

4. In your cluster page, click on the Applications tab and check your port for JupyterHub.



Application user interfaces Info

Applications installed on your Amazon EMR cluster publish user interfaces (UI) as websites. You can use these to monitor cluster activity.

● On-cluster application UIs
On-cluster UIs are only available while your cluster is running. Use the following links to get started. To access all the application UIs, set up SSH tunnelling.

○ Persistent application UIs
Persistent UIs don't require SSH tunnelling. They are hosted off the cluster and are available for 30 days after an application ends.

Live application UIs
These on-cluster application UIs are available without SSH tunnelling.

Application UIs
Spark history server UI

Application UIs on the primary node
These require SSH tunnelling to be enabled.

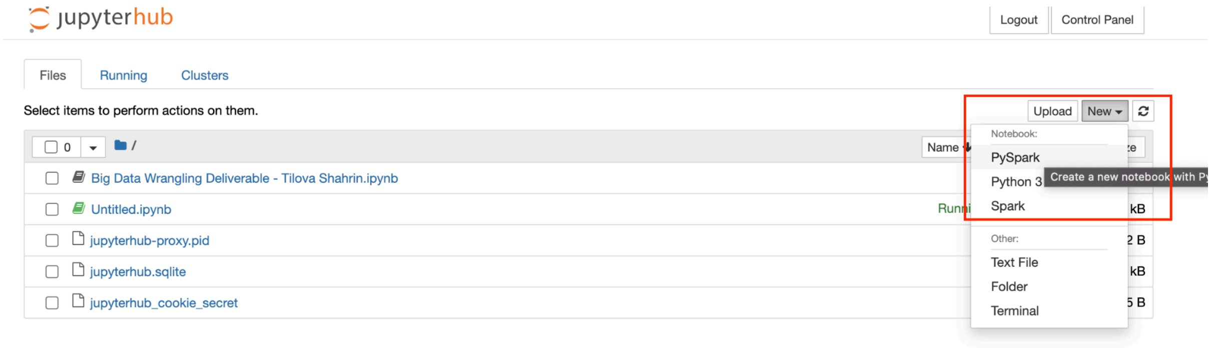
Application	UI URL
HDFS name node	http://ec2-18-118-27-195.us-east-2.compute.amazonaws.com:9870/
Hue	http://ec2-18-118-27-195.us-east-2.compute.amazonaws.com:8888/
JupyterHub	https://ec2-18-118-27-195.us-east-2.compute.amazonaws.com:9443/
Livy	http://ec2-18-118-27-195.us-east-2.compute.amazonaws.com:8998/

a. Type exit in your terminal and apply this command (follow pem file). *ssh -i aws_cloud_tilova.pem -L 9995:localhost:9443 hadoop@ec2-18-118-27-195.us-east-2.compute.amazonaws.com*

b. Make sure the port matches your jupyter hub port. Once it’s connected click on the link and proceed to ‘Advanced’ button and click on the link.

c. Once you enter the login page, use jovyan as username and jupyter as password.

d. When starting a new notebook, make sure it is a PySpark kernel.



Select items to perform actions on them.

Big Data Wrangling Deliverable - Tilova Shahrin.ipynb

Untitled.ipynb

jupyterhub-proxy.pid

jupyterhub.sqlite

jupyterhub_cookie_secret

e.

Appendix F

f. Continue onto the jupyter notebook.

5. On the terminal, start to merge the filtered data frame using this command.

a.

```
hadoop fs -getmerge hdfs://ec2-18-118-205-216.us-east-2.compute.amazonaws.com:8020/user/hadoop/eng_1M_1gram file_name_filtered_data.csv
```

b. Apply the csv file onto the s3 bucket using this command.

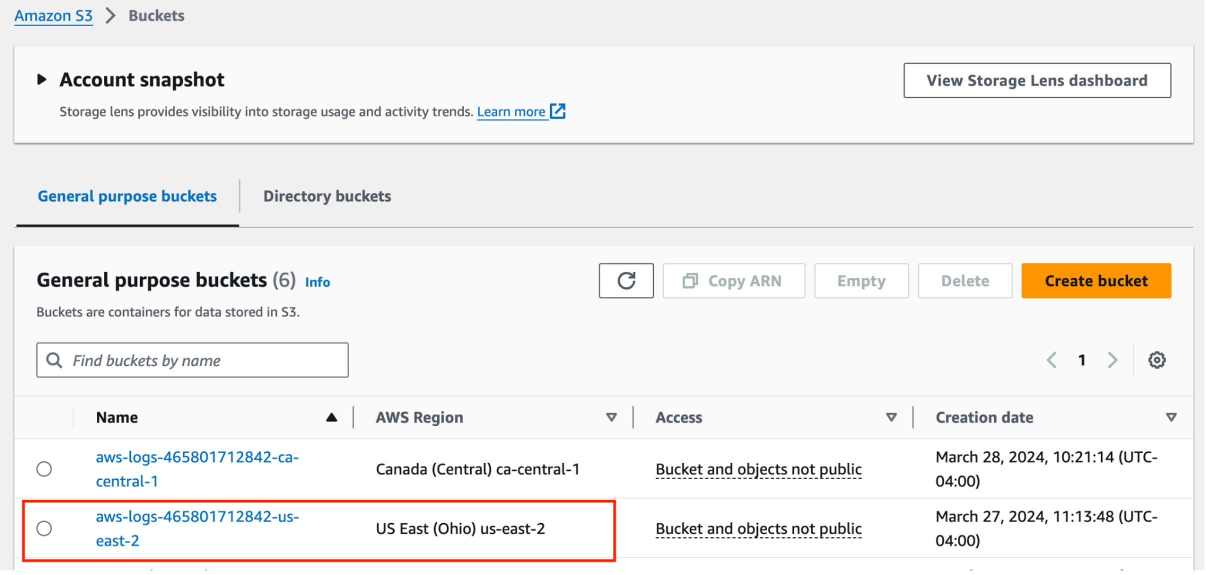
```
aws s3 mv file_name_filtered_data.csv s3://aws-logs-465801712842-us-east-2/eng_1M_1gram.csv
```

c. Then, check if the csv file is in the correct s3 bucket using this command.

```
aws s3 ls s3://aws-logs-465801712842-us-east-2
```

d. You should be able to see your csv file in your aws s3 bucket. Click on the appropriate bucket.

e.



Account snapshot

Storage lens provides visibility into storage usage and activity trends. Learn more

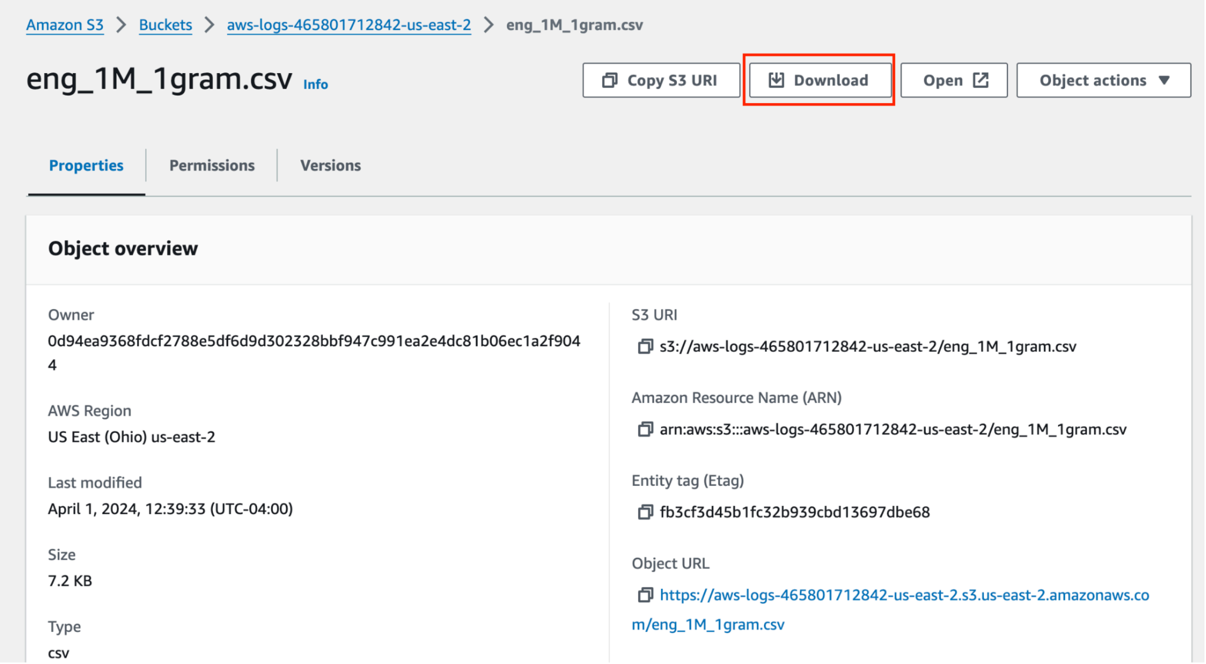
General purpose buckets (6) Info

Buckets are containers for data stored in S3.

	Name	AWS Region	Access	Creation date
○	aws-logs-465801712842-ca-central-1	Canada (Central) ca-central-1	Bucket and objects not public	March 28, 2024, 10:21:14 (UTC-04:00)
○	aws-logs-465801712842-us-east-2	US East (Ohio) us-east-2	Bucket and objects not public	March 27, 2024, 11:13:48 (UTC-04:00)

f. Then click on the file you've saved and click the download button to open your csv file.

g.



Amazon S3 > Buckets > aws-logs-465801712842-us-east-2 > eng_1M_1gram.csv
Copy S3 URI
Download
Open
Object actions
Properties
Permissions
Versions

eng_1M_1gram.csv Info

Object overview

Owner
0d94ea9368fdcf2788e5df6d9d302328bbf947c991ea2e4dc81b06ec1a2f9044

AWS Region
US East (Ohio) us-east-2

Last modified
April 1, 2024, 12:39:33 (UTC-04:00)

Size
7.2 KB

Type
csv

S3 URI
s3://aws-logs-465801712842-us-east-2/eng_1M_1gram.csv

Amazon Resource Name (ARN)
arn:aws:s3:::aws-logs-465801712842-us-east-2/eng_1M_1gram.csv

Entity tag (Etag)
fb3cf3d45b1fc32b939cbd13697dbe68

Object URL
https://aws-logs-465801712842-us-east-2.s3.us-east-2.amazonaws.com/eng_1M_1gram.csv

6. In the same page from the image shown above, grab your S3 URI and copy it so you can read csv from your local jupyter notebook. Check Jupyter Notebook - 'Big Data Local - Tilova Shahrin.ipynb'
7. Check Jupyter Notebook - 'Big Data Local - Tilova Shahrin.ipynb'
8. Hadoop integrates with external libraries to provide machine learning capabilities. Spark has built-in machine learning libraries. You would need Spark to run jupyter hub, which runs much faster than Hadoop. Hadoop runs at a lower cost since it relies on any disk storage type for data processing. Spark uses RAM for in memory processing so it costs more than Hadoop.