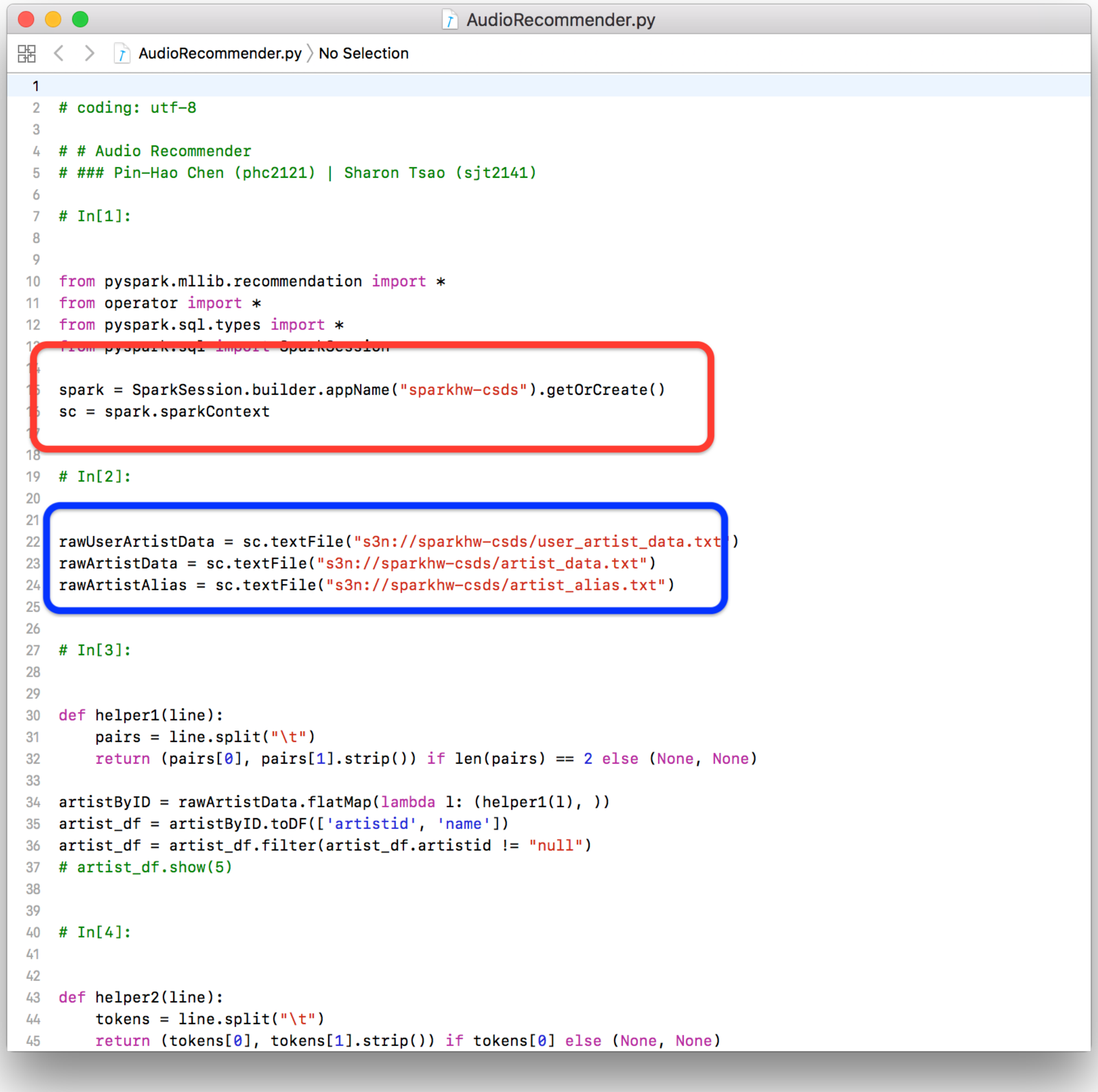
Pin-Hao Chen (phc2121)
Sharon Tsao (sjt2141)

Audio Recommender on AWS

1) We’re uploading the files to our bucket, sparkhw-csds

2) In order to run our script on AWS, we’ve added a few lines to the code to initialize SparkContext (sc). This includes importing and creating Sparksession, as shown in the image below in red. Additionally, we changed the path to the input file location of the bucket on AWS, as shown below in blue. The file names are in the format of “s3n://bucketname/filename”.

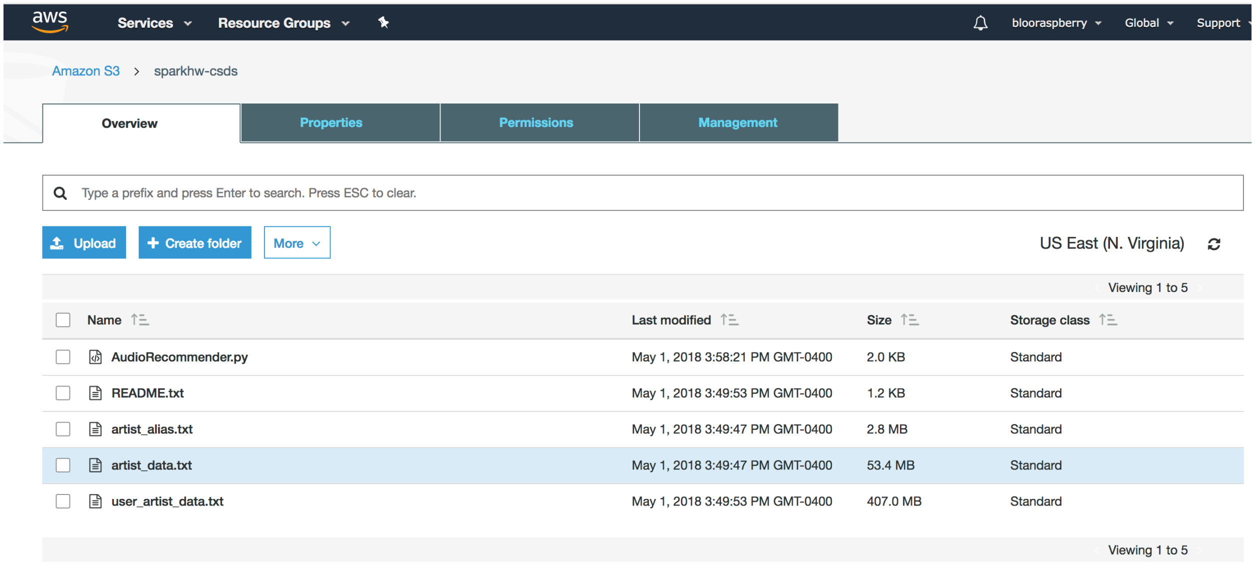


```

# coding: utf-8

# # Audio Recommender
# ### Pin-Hao Chen (phc2121) | Sharon Tsao (sjt2141)

# In[1]:


from pyspark.mllib.recommendation import *
from operator import *
from pyspark.sql.types import *
from pyspark.sql import SparkSession

spark = SparkSession.builder.appName("sparkhw-csds").getOrCreate()
sc = spark.sparkContext


# In[2]:


rawUserArtistData = sc.textFile("s3n://sparkhw-csds/user_artist_data.txt")
rawArtistData = sc.textFile("s3n://sparkhw-csds/artist_data.txt")
rawArtistAlias = sc.textFile("s3n://sparkhw-csds/artist_alias.txt")


# In[3]:


def helper1(line):
    pairs = line.split("\t")
    return (pairs[0], pairs[1].strip()) if len(pairs) == 2 else (None, None)

artistByID = rawArtistData.flatMap(lambda l: (helper1(l), ))
artist_df = artistByID.toDF(['artistid', 'name'])
artist_df = artist_df.filter(artist_df.artistid != "null")
# artist_df.show(5)


# In[4]:


def helper2(line):
    tokens = line.split("\t")
    return (tokens[0], tokens[1].strip()) if tokens[0] else (None, None)
```

Pin-Hao Chen (phc2121)
Sharon Tsao (sjt2141)

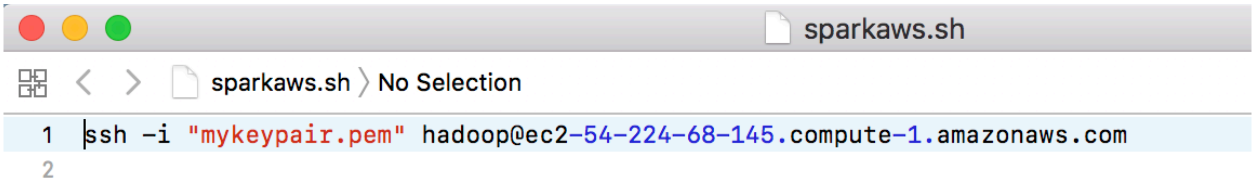
Audio Recommender on AWS

3) These are the files we've used in our bucket.



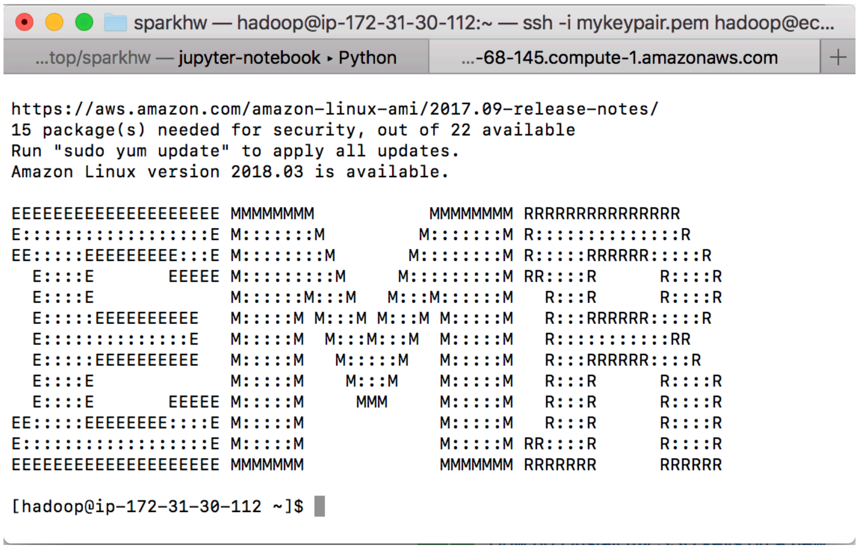
Name	Last modified	Size	Storage class
AudioRecommender.py	May 1, 2018 3:58:21 PM GMT-0400	2.0 KB	Standard
README.txt	May 1, 2018 3:49:53 PM GMT-0400	1.2 KB	Standard
artist_alias.txt	May 1, 2018 3:49:47 PM GMT-0400	2.8 MB	Standard
artist_data.txt	May 1, 2018 3:49:47 PM GMT-0400	53.4 MB	Standard
user_artist_data.txt	May 1, 2018 3:49:53 PM GMT-0400	407.0 MB	Standard

4) Here we are connecting to the instance through SSH by the permission key. We also changed the "mykeypair.pem" file to the 400 mode. Next we ran the following code in terminal.



```
ssh -i "mykeypair.pem" hadoop@ec2-54-224-68-145.compute-1.amazonaws.com
```

5) Screenshot of connecting to AWS EMR.

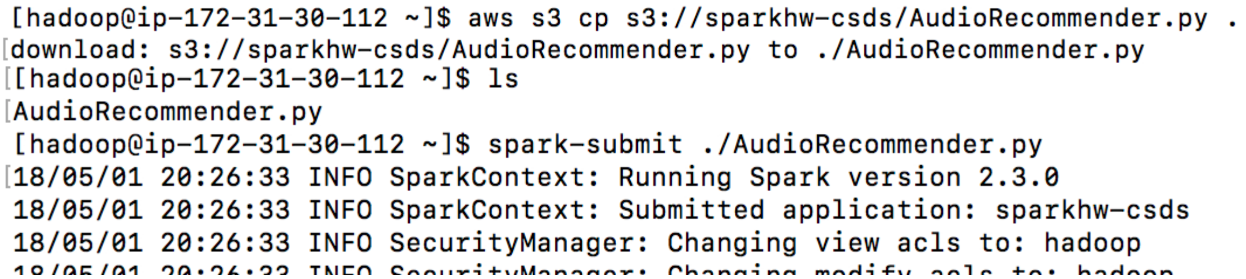

sparkhw — hadoop@ip-172-31-30-112:~ — ssh -i mykeypair.pem hadoop@ec...
...top/sparkhw — jupyter-notebook ▸ Python
...-68-145.compute-1.amazonaws.com

```
https://aws.amazon.com/amazon-linux-ami/2017.09-release-notes/
15 package(s) needed for security, out of 22 available
Run "sudo yum update" to apply all updates.
Amazon Linux version 2018.03 is available.

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-172-31-30-112 ~]$
```

Pin-Hao Chen (phc2121)
Sharon Tsao (sjt2141)

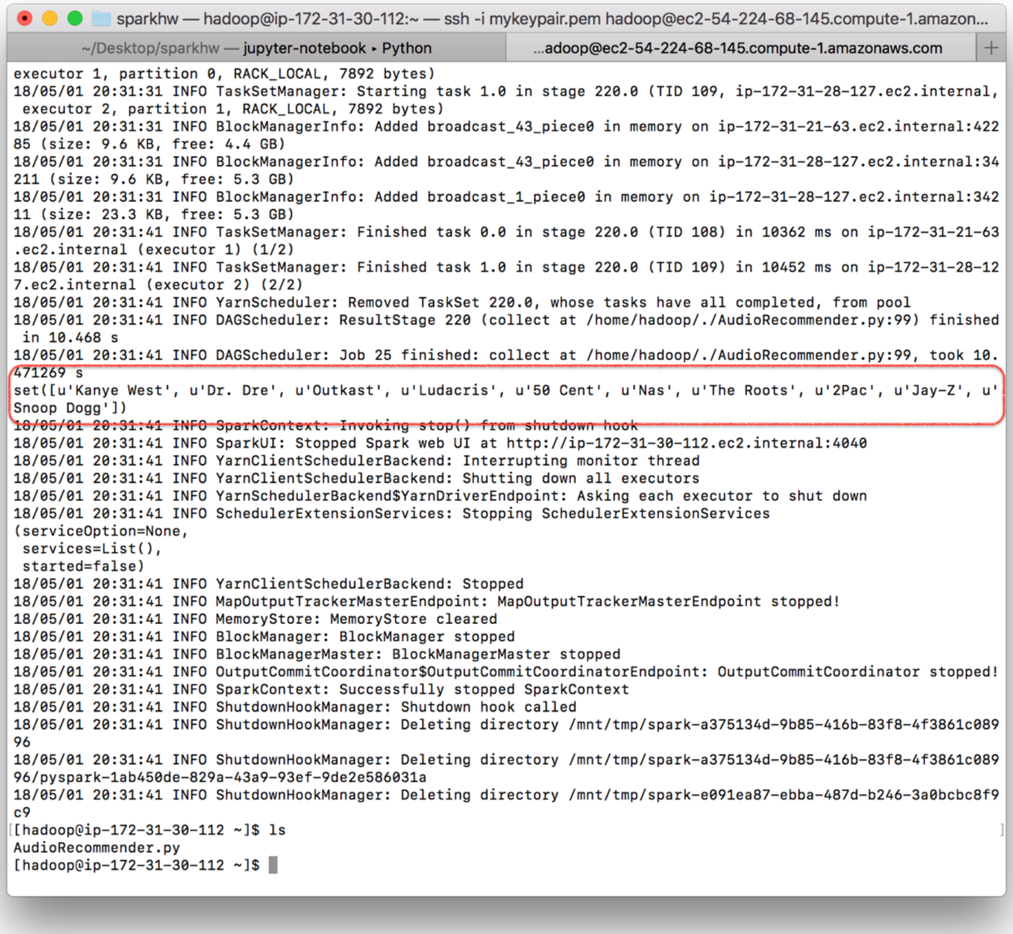
Audio Recommender on AWS

6) We copied over our python script and executed spark-submit.



```
 [hadoop@ip-172-31-30-112 ~]$ aws s3 cp s3://sparkhw-csds/AudioRecommender.py .
[download: s3://sparkhw-csds/AudioRecommender.py to ./AudioRecommender.py
[[hadoop@ip-172-31-30-112 ~]$ ls
[AudioRecommender.py
 [hadoop@ip-172-31-30-112 ~]$ spark-submit ./AudioRecommender.py
[18/05/01 20:26:33 INFO SparkContext: Running Spark version 2.3.0
 18/05/01 20:26:33 INFO SparkContext: Submitted application: sparkhw-csds
 18/05/01 20:26:33 INFO SecurityManager: Changing view acls to: hadoop
 18/05/01 20:26:33 INFO SecurityManager: Changing modify acls to: hadoop
```

7) The top 10 Audio Recommender artists are printed, as circled in red in the image below.


sparkhw — hadoop@ip-172-31-30-112:~ — ssh -i mykeypair.pem hadoop@ec2-54-224-68-145.compute-1.amazon...
~/Desktop/sparkhw — jupyter-notebook ▸ Python
...adoop@ec2-54-224-68-145.compute-1.amazonaws.com
+

```
executor 1, partition 0, RACK_LOCAL, 7892 bytes)
18/05/01 20:31:31 INFO TaskSetManager: Starting task 1.0 in stage 220.0 (TID 109, ip-172-31-28-127.ec2.internal,
 executor 2, partition 1, RACK_LOCAL, 7892 bytes)
18/05/01 20:31:31 INFO BlockManagerInfo: Added broadcast_43_piece0 in memory on ip-172-31-21-63.ec2.internal:422
85 (size: 9.6 KB, free: 4.4 GB)
18/05/01 20:31:31 INFO BlockManagerInfo: Added broadcast_43_piece0 in memory on ip-172-31-28-127.ec2.internal:34
211 (size: 9.6 KB, free: 5.3 GB)
18/05/01 20:31:31 INFO BlockManagerInfo: Added broadcast_1_piece0 in memory on ip-172-31-28-127.ec2.internal:342
11 (size: 23.3 KB, free: 5.3 GB)
18/05/01 20:31:41 INFO TaskSetManager: Finished task 0.0 in stage 220.0 (TID 108) in 10362 ms on ip-172-31-21-63
.ec2.internal (executor 1) (1/2)
18/05/01 20:31:41 INFO TaskSetManager: Finished task 1.0 in stage 220.0 (TID 109) in 10452 ms on ip-172-31-28-12
7.ec2.internal (executor 2) (2/2)
18/05/01 20:31:41 INFO YarnScheduler: Removed TaskSet 220.0, whose tasks have all completed, from pool
18/05/01 20:31:41 INFO DAGScheduler: ResultStage 220 (collect at /home/hadoop/./AudioRecommender.py:99) finished
 in 10.468 s
18/05/01 20:31:41 INFO DAGScheduler: Job 25 finished: collect at /home/hadoop/./AudioRecommender.py:99, took 10.
471269 s
set([u'Kanye West', u'Dr. Dre', u'Outkast', u'Ludacris', u'50 Cent', u'Nas', u'The Roots', u'2Pac', u'Jay-Z', u'
Snoop Dogg'])
18/05/01 20:31:41 INFO SparkContext: Invoking stop() from shutdown hook
18/05/01 20:31:41 INFO SparkUI: Stopped Spark web UI at http://ip-172-31-30-112.ec2.internal:4040
18/05/01 20:31:41 INFO YarnClientSchedulerBackend: Interrupting monitor thread
18/05/01 20:31:41 INFO YarnClientSchedulerBackend: Shutting down all executors
18/05/01 20:31:41 INFO YarnSchedulerBackend$YarnDriverEndpoint: Asking each executor to shut down
18/05/01 20:31:41 INFO SchedulerExtensionServices: Stopping SchedulerExtensionServices
(serviceOption=None,
 services=List(),
 started=false)
18/05/01 20:31:41 INFO YarnClientSchedulerBackend: Stopped
18/05/01 20:31:41 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
18/05/01 20:31:41 INFO MemoryStore: MemoryStore cleared
18/05/01 20:31:41 INFO BlockManager: BlockManager stopped
18/05/01 20:31:41 INFO BlockManagerMaster: BlockManagerMaster stopped
18/05/01 20:31:41 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
18/05/01 20:31:41 INFO SparkContext: Successfully stopped SparkContext
18/05/01 20:31:41 INFO ShutdownHookManager: Shutdown hook called
18/05/01 20:31:41 INFO ShutdownHookManager: Deleting directory /mnt/tmp/spark-a375134d-9b85-416b-83f8-4f3861c089
96
18/05/01 20:31:41 INFO ShutdownHookManager: Deleting directory /mnt/tmp/spark-a375134d-9b85-416b-83f8-4f3861c089
96/pyspark-1ab450de-829a-43a9-93ef-9de2e586031a
18/05/01 20:31:41 INFO ShutdownHookManager: Deleting directory /mnt/tmp/spark-e091ea87-ebba-487d-b246-3a0bcbc8f9
c9
[[hadoop@ip-172-31-30-112 ~]$ ls
AudioRecommender.py
[hadoop@ip-172-31-30-112 ~]$ 
```