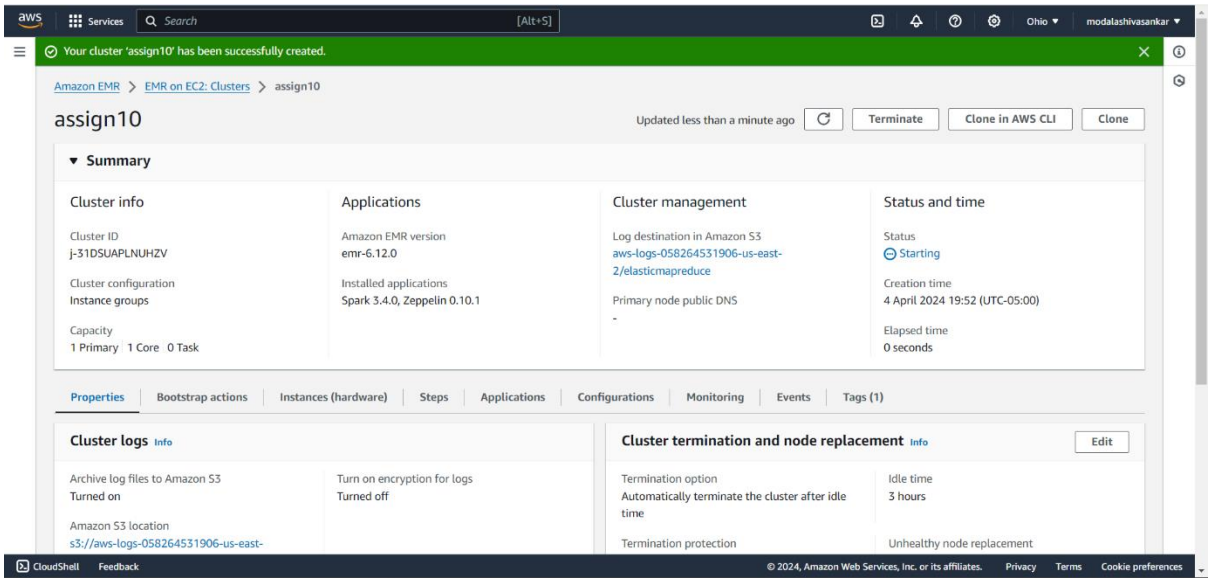
CSP 554 Big Data Technologies

Assignment – #10

Shiva Sankar Modala(A20517528)

1) Start up a EMR cluster as previously, but instead of choosing the “Core Hadoop” configuration chose the “Spark” configuration (see below), otherwise proceed as before.

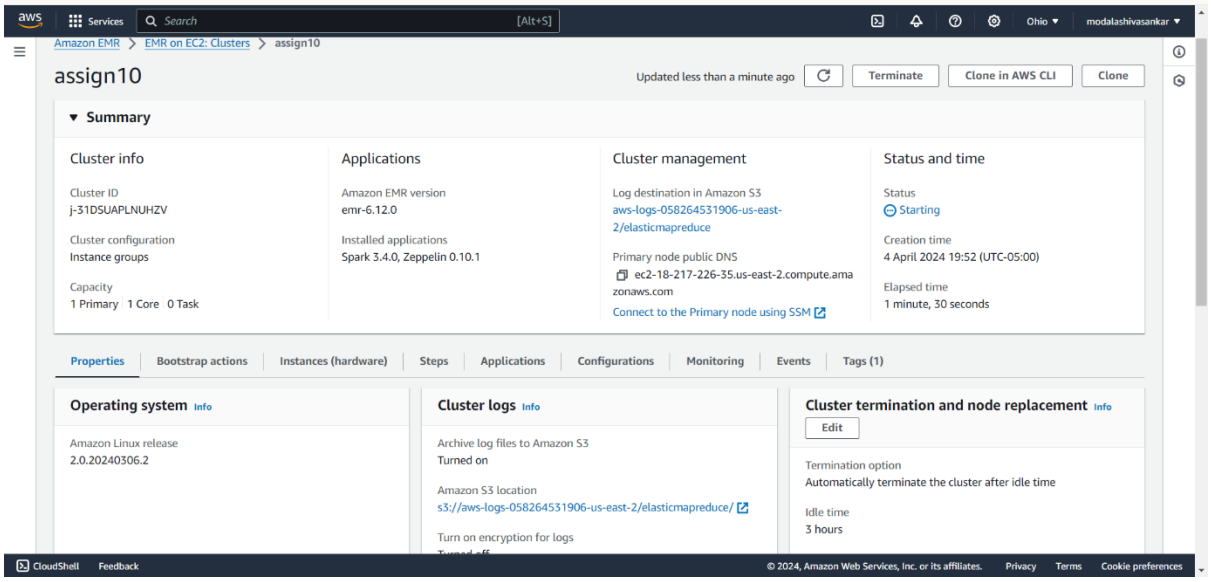


2) At a later point in these instructions, you will need to use the public DNS name of the primary (master) node of your EMR cluster. To retrieve it using the Amazon EMR console

a) Find the EMR main page.

b) On the Clusters menu selection, select the link for your cluster.

c) Note the Primary public DNS value that appears at the top of the cluster details page.



3) Download consume.py and log4j.properties files from the assignment to your local PC or MAC

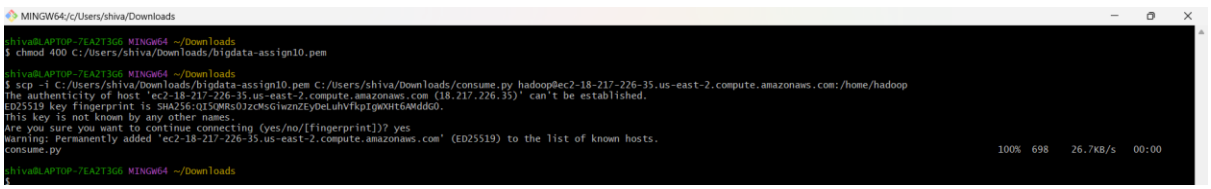
4) There is one item you must change in consume.py. In the following line you must replace

<Master public DNS> with your own public DNS name (found as described above)

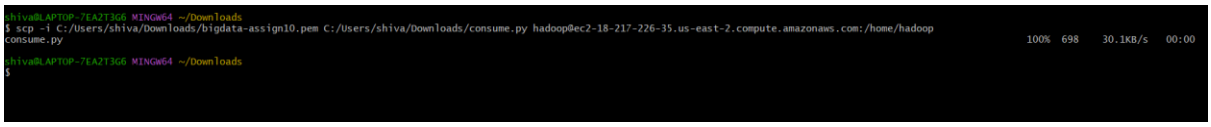
consume.py

```
from pyspark import SparkContext
from pyspark.streaming import StreamingContext

# Create a local StreamingContext with a batch interval of 10 seconds
sc = SparkContext("yarn", "NetworkWordCount")
ssc = StreamingContext(sc, 10)

# Create a DStream
lines = ssc.socketTextStream("ec2-18-217-226-35.us-east-2.compute.amazonaws.com", 3333)

# Split each line into words
words = lines.flatMap(lambda line: line.split(" "))

# Count each word in each batch
pairs = words.map(lambda word: (word, 1))
wordCounts = pairs.reduceByKey(lambda x, y: x + y)

# Print each batch
wordCounts.pprint()

ssc.start()             # Start the computation
ssc.awaitTermination()  # Wait for the computation to terminate


```

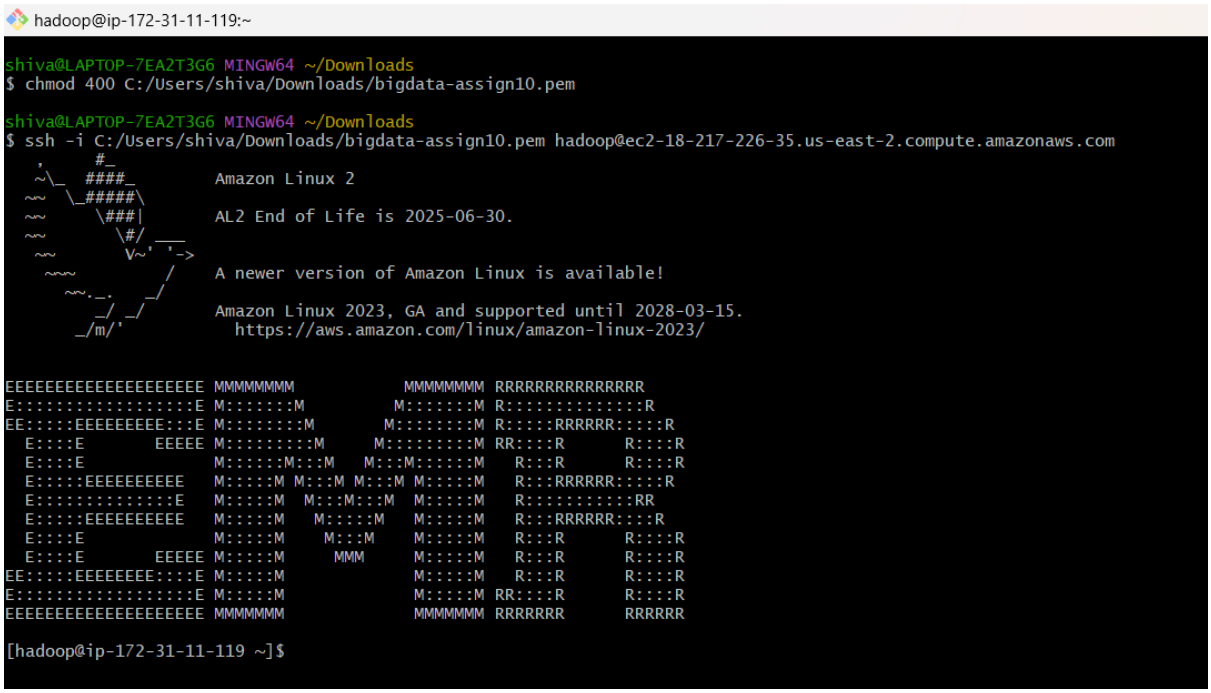
5) scp this modified consume.py file to your EMR cluster primary (master) node. You may need to answer a security question with “Y/y” or “Yes”.

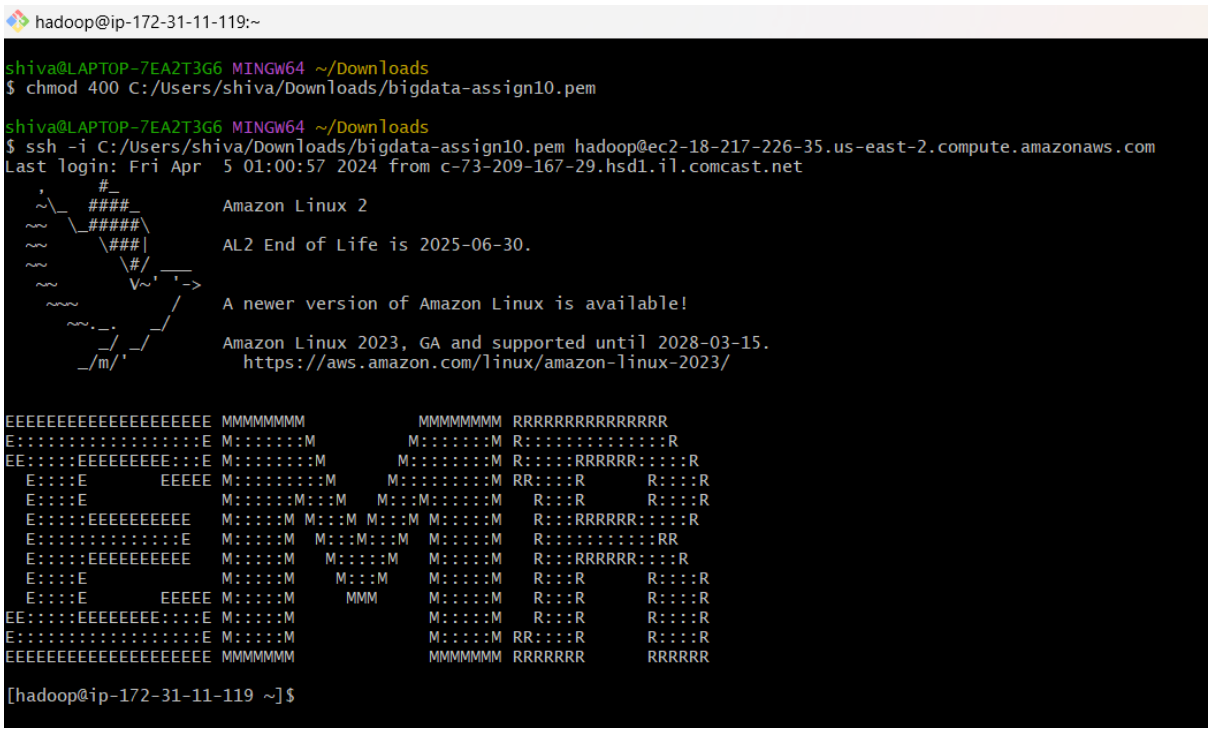


6) Then scp the file log4j.properties to your EMR cluster primary (master) node.



7) Open two terminal sessions to the EMR primary node. We will call one the EC2-1 window and the other the EC2-2 window.

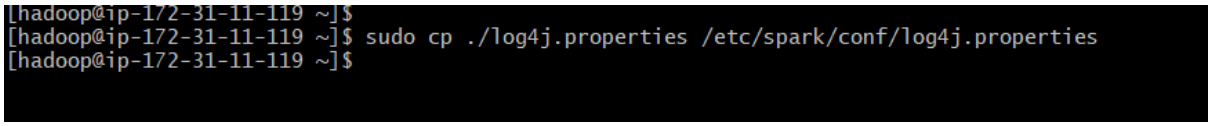




8) In the EC2-1 window enter the following:

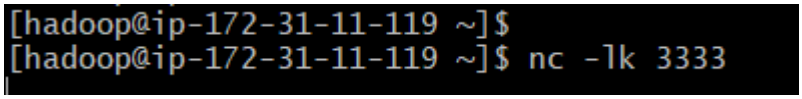
sudo cp ./log4j.properties /etc/spark/conf/log4j.properties

This changes the logging properties to turn off “INFO” messages to allow easier viewing of the results of the stream processing job. But it is not something you always want to disable.



9) In the EC2-1 window enter the following command to open a TCP (socket) connection on port 3333

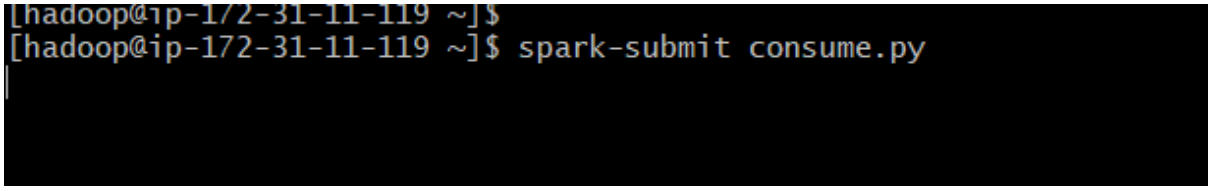
nc -lk 3333



10) In the EC2-2 window enter the following command:

spark-submit consume.py

This takes a while to start up. So, wait for some messages issued to the console before continuing. Note, when you do this you might see a message beginning with “WARN StreamingContext:...” which you can ignore.



```
hadoop@ip-172-31-11-119:~
24/04/05 01:15:10 INFO BlockManagerInfo: Removed broadcast_15_piece0 on ip-172-31-11-119.us-east-2.compute.internal:45633 in memory (size: 6.1 KiB, free: 912.2 MiB)
24/04/05 01:15:10 INFO BlockManagerInfo: Removed broadcast_15_piece0 on ip-172-31-15-199.us-east-2.compute.internal:33241 in memory (size: 6.1 KiB, free: 4.8 GiB)
24/04/05 01:15:10 INFO BlockManagerInfo: Removed broadcast_21_piece0 on ip-172-31-11-119.us-east-2.compute.internal:45633 in memory (size: 6.1 KiB, free: 912.2 MiB)
24/04/05 01:15:10 INFO BlockManagerInfo: Removed broadcast_21_piece0 on ip-172-31-15-199.us-east-2.compute.internal:33241 in memory (size: 6.1 KiB, free: 4.8 GiB)
24/04/05 01:15:10 INFO SparkContext: Starting job: runJob at PythonRDD.scala:179
24/04/05 01:15:10 INFO DAGScheduler: Got job 27 (runJob at PythonRDD.scala:179) with 1 output partitions
24/04/05 01:15:10 INFO DAGScheduler: Final stage: ResultStage 54 (runJob at PythonRDD.scala:179)
24/04/05 01:15:10 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 53)
24/04/05 01:15:10 INFO DAGScheduler: Missing parents: List()
24/04/05 01:15:10 INFO DAGScheduler: Submitting ResultStage 54 (PythonRDD[107] at RDD at PythonRDD.scala:53), which has no missing parents
24/04/05 01:15:10 INFO MemoryStore: Block broadcast_29 stored as values in memory (estimated size 11.2 KiB, free 912.0 MiB)
24/04/05 01:15:10 INFO MemoryStore: Block broadcast_29_piece0 stored as bytes in memory (estimated size 6.1 KiB, free 912.0 MiB)
24/04/05 01:15:10 INFO BlockManagerInfo: Added broadcast_29_piece0 in memory on ip-172-31-11-119.us-east-2.compute.internal:45633 (size: 6.1 KiB, free: 912.2 MiB)
24/04/05 01:15:10 INFO SparkContext: Created broadcast 29 from broadcast at DAGScheduler.scala:1592
24/04/05 01:15:10 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 54 (PythonRDD[107] at RDD at PythonRDD.scala:53) (first 15 tasks are for partitions Vector(1))
24/04/05 01:15:10 INFO YarnScheduler: Adding task set 54.0 with 1 tasks resource profile 0
24/04/05 01:15:10 INFO TaskSetManager: Starting task 0.0 in stage 54.0 (TID 97) (ip-172-31-15-199.us-east-2.compute.internal, executor 1, partition 1, PROCESS_LOCAL, 7192 bytes)
24/04/05 01:15:10 INFO BlockManagerInfo: Removed broadcast_13_piece0 on ip-172-31-11-119.us-east-2.compute.internal:45633 in memory (size: 6.1 KiB, free: 912.2 MiB)
24/04/05 01:15:10 INFO BlockManagerInfo: Removed broadcast_13_piece0 on ip-172-31-15-199.us-east-2.compute.internal:33241 in memory (size: 6.1 KiB, free: 4.8 GiB)
24/04/05 01:15:10 INFO BlockManagerInfo: Added broadcast_29_piece0 in memory on ip-172-31-15-199.us-east-2.compute.internal:33241 (size: 6.1 KiB, free: 4.8 GiB)
24/04/05 01:15:10 INFO BlockManagerInfo: Removed broadcast_19_piece0 on ip-172-31-11-119.us-east-2.compute.internal:45633 in memory (size: 6.1 KiB, free: 912.2 MiB)
24/04/05 01:15:10 INFO BlockManagerInfo: Removed broadcast_19_piece0 on ip-172-31-15-199.us-east-2.compute.internal:33241 in memory (size: 6.1 KiB, free: 4.8 GiB)
24/04/05 01:15:10 INFO BlockManagerInfo: Removed broadcast_20_piece0 on ip-172-31-11-119.us-east-2.compute.internal:45633 in memory (size: 6.1 KiB, free: 912.2 MiB)
24/04/05 01:15:10 INFO BlockManagerInfo: Removed broadcast_20_piece0 on ip-172-31-15-199.us-east-2.compute.internal:33241 in memory (size: 6.1 KiB, free: 4.8 GiB)
24/04/05 01:15:10 INFO BlockManagerInfo: Removed broadcast_24_piece0 on ip-172-31-15-199.us-east-2.compute.internal:33241 in memory (size: 6.1 KiB, free: 4.8 GiB)
24/04/05 01:15:10 INFO BlockManagerInfo: Removed broadcast_24_piece0 on ip-172-31-11-119.us-east-2.compute.internal:45633 in memory (size: 6.1 KiB, free: 912.2 MiB)
24/04/05 01:15:10 INFO BlockManagerInfo: Removed broadcast_26_piece0 on ip-172-31-15-199.us-east-2.compute.internal:33241 in memory (size: 6.1 KiB, free: 4.8 GiB)
24/04/05 01:15:10 INFO BlockManagerInfo: Removed broadcast_26_piece0 on ip-172-31-11-119.us-east-2.compute.internal:45633 in memory (size: 6.1 KiB, free: 912.2 MiB)
24/04/05 01:15:10 INFO BlockManagerInfo: Removed broadcast_14_piece0 on ip-172-31-11-119.us-east-2.compute.internal:45633 in memory (size: 6.1 KiB, free: 912.2 MiB)
24/04/05 01:15:10 INFO BlockManagerInfo: Removed broadcast_14_piece0 on ip-172-31-15-199.us-east-2.compute.internal:33241 in memory (size: 6.1 KiB, free: 4.8 GiB)
24/04/05 01:15:10 INFO BlockManagerInfo: Removed broadcast_10_piece0 on ip-172-31-11-119.us-east-2.compute.internal:45633 in memory (size: 6.1 KiB, free: 912.2 MiB)
24/04/05 01:15:10 INFO BlockManagerInfo: Removed broadcast_10_piece0 on ip-172-31-15-199.us-east-2.compute.internal:33241 in memory (size: 6.1 KiB, free: 4.8 GiB)
24/04/05 01:15:10 INFO TaskSetManager: Finished task 0.0 in stage 54.0 (TID 97) in 70 ms on ip-172-31-15-199.us-east-2.compute.internal (executor 1) (1/1)
24/04/05 01:15:10 INFO YarnScheduler: Removed TaskSet 54.0, whose tasks have all completed, from pool
24/04/05 01:15:10 INFO DAGScheduler: ResultStage 54 (runJob at PythonRDD.scala:179) finished in 0.078 s
24/04/05 01:15:10 INFO DAGScheduler: Job 27 is finished. Cancelling potential speculative or zombie tasks for this job
24/04/05 01:15:10 INFO YarnScheduler: Killing all running tasks in stage 54: Stage finished
24/04/05 01:15:10 INFO DAGScheduler: Job 27 finished: runJob at PythonRDD.scala:179, took 0.081552 s
-------------------------------------------
```

11) Now in the EC2-1 window enter one or more lines of text and press Enter/Return after each one including the last. You should see the word count results scroll by in the EC2-2 window

```
[hadoop@ip-172-31-11-119 ~]$ nc -lk 3333
Hi this is shiva
Big data refers to extremely large and diverse collections of structured, unstructured, and semi-structured data that continues to grow exponentially over time.
These datasets are so huge and complex in volume, velocity, and variety, that traditional data management systems cannot store, process, and analyze them.
```

```
-------------------------------------------
Time: 2024-04-05 01:15:20
-------------------------------------------
('Hi', 1)
('is', 1)
('shiva', 1)
('this', 1)
```

```
-------------------------------------------
Time: 2024-04-05 01:15:50
-------------------------------------------
('refers', 1)
('large', 1)
('diverse', 1)
('collections', 1)
('of', 1)
('unstructured,', 1)
('continues', 1)
('Big', 1)
('data', 2)
('to', 2)
...
```

```
-------------------------------------------
Time: 2024-04-05 01:16:30
-------------------------------------------
('These', 1)
('datasets', 1)
('are', 1)
('huge', 1)
('in', 1)
('volume,', 1)
('velocity,', 1)
('variety,', 1)
('traditional', 1)
('management', 1)
...
```

```
-------------------------------------------
Time: 2024-04-05 01:25:40
-------------------------------------------

^Z
[1]+  Stopped                 spark-submit consume.py
```

12) Remember to terminate your EMR instance after you are done!