


CSP 554

Assignment 10

Ashmita Gupta (A20512498)

1) Setting up EMR cluster with Spark configurations:

My cluster

Updated less than a minute ago 

▼ Summary

Cluster info

Cluster ID
j-2X2C0LO9AMDXX

Cluster configuration
Instance groups

Capacity
1 Primary | 1 Core | 0 Task



Applications


Amazon EMR version
emr-6.12.0

Installed applications
Spark 3.4.0, Zeppelin 0.10.1

Cluster management

Log destination in Amazon S3
[aws-logs-020428218454-us-east-1/elasticmapreduce](#)

Persistent application UIs
[Spark History Server](#) 
[YARN timeline server](#) 

Primary node public DNS
 [ec2-3-239-110-85.compute-1.amazonaws.com](#)
[Connect to the Primary node using SSH](#)

2) Modified consume.py with above DNS name:

```
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("ec2-3-239-110-85.compute-1.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
```

3) scp consume.py and log4j.properties:

```
gashm@Ashmita MINGW64 ~
$ cd OneDrive/Desktop/Big Data

gashm@Ashmita MINGW64 ~/OneDrive/Desktop/Big Data
$ scp -i emr-key-pair-2.pem consume.py hadoop@ec2-3-239-110-85.compute-1.amazona
ws.com:/home/hadoop
The authenticity of host 'ec2-3-239-110-85.compute-1.amazonaws.com (3.239.110.85
)' can't be established.
ED25519 key fingerprint is SHA256:S46pASyH00zfQ81S1SRKIdEqf0B4AUtPNSfL+8NvZvI.
This key is not known by any other names.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'ec2-3-239-110-85.compute-1.amazonaws.com' (ED25519)
to the list of known hosts.
consume.py                                100% 689   18.1KB/s   00:00

gashm@Ashmita MINGW64 ~/OneDrive/Desktop/Big Data
$ scp -i emr-key-pair-2.pem log4j.properties hadoop@ec2-3-239-110-85.compute-1.amazonaws.com:/home/hadoop
log4j.properties                          100% 3199   82.9KB/s   00:00

gashm@Ashmita MINGW64 ~/OneDrive/Desktop/Big Data
$ |
```

4) EC2-1 window the following commands/actions:

- sudo cp ./log4j.properties /etc/spark/conf/log4j.properties
- nc -lk 3333
- EC2-1 window enter one or more lines of text

```
gashm@Ashmita MINGW64 ~/OneDrive/Desktop/Big Data
$ ssh -i emr-key-pair-2.pem hadoop@ec2-3-239-110-85.compute-1.amazonaws.com

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

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

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

[hadoop@ip-172-31-12-40 ~]$ sudo cp ./log4j.properties /etc/spark/conf/log4j.properties
[hadoop@ip-172-31-12-40 ~]$ nc -lk 3333
bid data
big data big data big data
Ashmita Gupta Ashmita Gupta Ashmita
Ashmita Ashmita ashmita
```



```
23/11/14 00:35:40 INFO DAGScheduler: Final stage: ResultStage 102 (runJob at PythonRDD.scala:179)
23/11/14 00:35:40 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 101)
23/11/14 00:35:40 INFO DAGScheduler: Missing parents: List()
23/11/14 00:35:40 INFO DAGScheduler: Submitting ResultStage 102 (PythonRDD[203] at RDD at PythonRDD
23/11/14 00:35:40 INFO MemoryStore: Block broadcast_53 stored as values in memory (estimated size 1
23/11/14 00:35:40 INFO MemoryStore: Block broadcast_53_piece0 stored as bytes in memory (estimated
23/11/14 00:35:40 INFO BlockManagerInfo: Added broadcast_53_piece0 in memory on ip-172-31-12-40.ec2
23/11/14 00:35:40 INFO SparkContext: Created broadcast 53 from broadcast at DAGScheduler.scala:1592
23/11/14 00:35:40 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 102 (PythonRDD[203
23/11/14 00:35:40 INFO YarnScheduler: Adding task set 102.0 with 1 tasks resource profile 0
23/11/14 00:35:40 INFO TaskSetManager: Starting task 0.0 in stage 102.0 (TID 121) (ip-172-31-14-226
23/11/14 00:35:40 INFO BlockManagerInfo: Added broadcast_53_piece0 in memory on ip-172-31-14-226.ec
23/11/14 00:35:40 INFO TaskSetManager: Finished task 0.0 in stage 102.0 (TID 121) in 72 ms on ip-17
23/11/14 00:35:40 INFO YarnScheduler: Removed TaskSet 102.0, whose tasks have all completed, from p
23/11/14 00:35:40 INFO DAGScheduler: ResultStage 102 (runJob at PythonRDD.scala:179) finished in 0.
23/11/14 00:35:40 INFO DAGScheduler: Job 51 is finished. Cancelling potential speculative or zombie
23/11/14 00:35:40 INFO YarnScheduler: Killing all running tasks in stage 102: Stage finished
23/11/14 00:35:40 INFO DAGScheduler: Job 51 finished: runJob at PythonRDD.scala:179, took 0.089220
```

Time: 2023-11-14 00:35:40

('bid', 1)
('data', 1)

```
23/11/14 00:35:40 INFO JobScheduler: Finished job streaming job 1699922140000 ms.0 from job set of
23/11/14 00:35:40 INFO JobScheduler: Total delay: 0.383 s for time 1699922140000 ms (execution: 0.3
23/11/14 00:35:40 INFO PythonRDD: Removing RDD 193 from persistence list
23/11/14 00:35:40 INFO BlockManager: Removing RDD 193
23/11/14 00:35:40 INFO BlockRDD: Removing RDD 188 from persistence list
23/11/14 00:35:40 INFO SocketInputDStream: Removing blocks of RDD BlockRDD[188] at socketTextStream
23/11/14 00:35:40 INFO ReceivedBlockTracker: Deleting batches: 1699922120000 ms
23/11/14 00:35:40 INFO InputInfoTracker: remove old batch metadata: 1699922120000 ms
23/11/14 00:35:40 INFO BlockManager: Removing RDD 188
23/11/14 00:35:50 INFO JobScheduler: Added jobs for time 1699922150000 ms
23/11/14 00:35:50 INFO JobScheduler: Starting job streaming job 1699922150000 ms.0 from job set of
23/11/14 00:35:50 INFO SparkContext: Starting job: runJob at PythonRDD.scala:179
23/11/14 00:35:50 INFO DAGScheduler: Registering RDD 206 (call at /usr/lib/spark/python/lib/py4j-0.
23/11/14 00:35:50 INFO DAGScheduler: Got job 52 (runJob at PythonRDD.scala:179) with 1 output parti
23/11/14 00:35:50 INFO DAGScheduler: Final stage: ResultStage 104 (runJob at PythonRDD.scala:179)
```

```
23/11/14 00:36:00 INFO BlockManagerInfo: Added broadcast_58_piece0 in memory on ip-172-31-12-40.ec2.inte
23/11/14 00:36:00 INFO SparkContext: Created broadcast 58 from broadcast at DAGScheduler.scala:1592
23/11/14 00:36:00 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 110 (PythonRDD[219] at R
23/11/14 00:36:00 INFO YarnScheduler: Adding task set 110.0 with 1 tasks resource profile 0
23/11/14 00:36:00 INFO TaskSetManager: Starting task 0.0 in stage 110.0 (TID 126) (ip-172-31-14-226.ec2.1
23/11/14 00:36:00 INFO BlockManagerInfo: Added broadcast_58_piece0 in memory on ip-172-31-14-226.ec2.inte
23/11/14 00:36:00 INFO TaskSetManager: Finished task 0.0 in stage 110.0 (TID 126) in 59 ms on ip-172-31-1
23/11/14 00:36:00 INFO YarnScheduler: Removed TaskSet 110.0, whose tasks have all completed, from pool
23/11/14 00:36:00 INFO DAGScheduler: ResultStage 110 (runJob at PythonRDD.scala:179) finished in 0.067 s
23/11/14 00:36:00 INFO DAGScheduler: Job 55 is finished. Cancelling potential speculative or zombie tasks
23/11/14 00:36:00 INFO YarnScheduler: Killing all running tasks in stage 110: Stage finished
23/11/14 00:36:00 INFO DAGScheduler: Job 55 finished: runJob at PythonRDD.scala:179, took 0.070478 s
```

Time: 2023-11-14 00:36:00

('big', 3)
('data', 3)

```
23/11/14 00:36:00 INFO JobScheduler: Finished job streaming job 1699922160000 ms.0 from job set of time 1
23/11/14 00:36:00 INFO JobScheduler: Total delay: 0.306 s for time 1699922160000 ms (execution: 0.271 s)
23/11/14 00:36:00 INFO PythonRDD: Removing RDD 209 from persistence list
23/11/14 00:36:00 INFO BlockRDD: Removing RDD 204 from persistence list
23/11/14 00:36:00 INFO BlockManager: Removing RDD 209
23/11/14 00:36:00 INFO SocketInputDStream: Removing blocks of RDD BlockRDD[204] at socketTextStream at Na
23/11/14 00:36:00 INFO ReceivedBlockTracker: Deleting batches: 1699922140000 ms
23/11/14 00:36:00 INFO InputInfoTracker: remove old batch metadata: 1699922140000 ms
23/11/14 00:36:00 INFO BlockManager: Removing RDD 204
23/11/14 00:36:10 INFO JobScheduler: Added jobs for time 1699922170000 ms
23/11/14 00:36:10 INFO JobScheduler: Starting job streaming job 1699922170000 ms.0 from job set of time 1
23/11/14 00:36:10 INFO SparkContext: Starting job: runJob at PythonRDD.scala:179
23/11/14 00:36:10 INFO DAGScheduler: Registering RDD 222 (call at /usr/lib/spark/python/lib/py4j-0.10.9.7
23/11/14 00:36:10 INFO DAGScheduler: Got job 56 (runJob at PythonRDD.scala:179) with 1 output partitions
```



```
23/11/14 00:36:30 INFO YarnScheduler: Adding task set 122.0 with 1 tasks resource profile 0
23/11/14 00:36:30 INFO TaskSetManager: Starting task 0.0 in stage 122.0 (TID 133) (ip-172-31-14-2
23/11/14 00:36:30 INFO BlockManagerInfo: Added broadcast_65_piece0 in memory on ip-172-31-14-226.
23/11/14 00:36:30 INFO TaskSetManager: Finished task 0.0 in stage 122.0 (TID 133) in 65 ms on ip-
23/11/14 00:36:30 INFO YarnScheduler: Removed TaskSet 122.0, whose tasks have all completed, from
23/11/14 00:36:30 INFO DAGScheduler: ResultStage 122 (runJob at PythonRDD.scala:179) finished in
23/11/14 00:36:30 INFO DAGScheduler: Job 61 is finished. Cancelling potential speculative or zomb
23/11/14 00:36:30 INFO YarnScheduler: Killing all running tasks in stage 122: Stage finished
23/11/14 00:36:30 INFO DAGScheduler: Job 61 finished: runJob at PythonRDD.scala:179, took 0.07282
```

Time: 2023-11-14 00:36:30

```
-----
('', 1)
('Ashmita', 3)
('Gupta', 2)
```

```
23/11/14 00:36:30 INFO JobScheduler: Finished job streaming job 1699922190000 ms.0 from job set c
23/11/14 00:36:30 INFO JobScheduler: Total delay: 0.284 s for time 1699922190000 ms (execution: 0
23/11/14 00:36:30 INFO PythonRDD: Removing RDD 233 from persistence list
23/11/14 00:36:30 INFO BlockManager: Removing RDD 233
23/11/14 00:36:30 INFO BlockRDD: Removing RDD 228 from persistence list
23/11/14 00:36:30 INFO SocketInputStream: Removing blocks of RDD BlockRDD[228] at socketTextStre
23/11/14 00:36:30 INFO ReceivedBlockTracker: Deleting batches: 1699922170000 ms
23/11/14 00:36:30 INFO InputInfoTracker: remove old batch metadata: 1699922170000 ms
23/11/14 00:36:30 INFO BlockManager: Removing RDD 228
23/11/14 00:36:40 INFO JobScheduler: Added jobs for time 1699922200000 ms
23/11/14 00:36:40 INFO JobScheduler: Starting job streaming job 1699922200000 ms.0 from job set c
23/11/14 00:36:40 INFO SparkContext: Starting job: runJob at PythonRDD.scala:179
23/11/14 00:36:40 INFO DAGScheduler: Registering RDD 246 (call at /usr/lib/spark/python/lib/py4j-
23/11/14 00:36:40 INFO DAGScheduler: Got job 62 (runJob at PythonRDD.scala:179) with 1 output par
23/11/14 00:36:40 INFO DAGScheduler: Final stage: ResultStage 124 (runJob at PythonRDD.scala:179)
```

```
23/11/14 00:36:50 INFO SparkContext: Created broadcast 70 from broadcast at DAGSche
23/11/14 00:36:50 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 13
23/11/14 00:36:50 INFO YarnScheduler: Adding task set 130.0 with 1 tasks resource p
23/11/14 00:36:50 INFO TaskSetManager: Starting task 0.0 in stage 130.0 (TID 138)
23/11/14 00:36:50 INFO BlockManagerInfo: Added broadcast_70_piece0 in memory on ip-
23/11/14 00:36:50 INFO TaskSetManager: Finished task 0.0 in stage 130.0 (TID 138)
23/11/14 00:36:50 INFO YarnScheduler: Removed TaskSet 130.0, whose tasks have all c
23/11/14 00:36:50 INFO DAGScheduler: ResultStage 130 (runJob at PythonRDD.scala:179)
23/11/14 00:36:50 INFO DAGScheduler: Job 65 is finished. Cancelling potential specu
23/11/14 00:36:50 INFO YarnScheduler: Killing all running tasks in stage 130: Stage
23/11/14 00:36:50 INFO DAGScheduler: Job 65 finished: runJob at PythonRDD.scala:179
```

Time: 2023-11-14 00:36:50

```
-----
('Ashmita', 2)
('ashmita', 1)
```

```
23/11/14 00:36:50 INFO JobScheduler: Finished job streaming job 1699922210000 ms.0
23/11/14 00:36:50 INFO JobScheduler: Total delay: 0.295 s for time 1699922210000 ms
23/11/14 00:36:50 INFO PythonRDD: Removing RDD 249 from persistence list
23/11/14 00:36:50 INFO BlockRDD: Removing RDD 244 from persistence list
23/11/14 00:36:50 INFO BlockManager: Removing RDD 249
23/11/14 00:36:50 INFO SocketInputStream: Removing blocks of RDD BlockRDD[244] at
23/11/14 00:36:50 INFO ReceivedBlockTracker: Deleting batches: 1699922190000 ms
23/11/14 00:36:50 INFO InputInfoTracker: remove old batch metadata: 1699922190000
23/11/14 00:36:50 INFO BlockManager: Removing RDD 244
23/11/14 00:37:00 INFO JobScheduler: Added jobs for time 1699922220000 ms
23/11/14 00:37:00 INFO JobScheduler: Starting job streaming job 1699922220000 ms.0
23/11/14 00:37:00 INFO SparkContext: Starting job: runJob at PythonRDD.scala:179
23/11/14 00:37:00 INFO DAGScheduler: Registering RDD 262 (call at /usr/lib/spark/py
23/11/14 00:37:00 INFO DAGScheduler: Got job 66 (runJob at PythonRDD.scala:179) wi
23/11/14 00:37:00 INFO DAGScheduler: Final stage: ResultStage 132 (runJob at Python
23/11/14 00:37:00 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage
23/11/14 00:37:00 INFO DAGScheduler: Missing parents: List()
23/11/14 00:37:00 INFO DAGScheduler: Submitting ResultStage 132 (PythonRDD[266] at
23/11/14 00:37:00 INFO MemoryStore: Block broadcast_71 stored as values in memory
```