Rishabh Jain || A20495530 || rjain35@hawk.iit.edu

Assignment 10 CSP 554

Performing SSH to the cluster

```
➜  Desktop  ssh -i ~/Desktop/new-key-pair-emr.pem hadoop@ec2-44-203-246-127.compute-1.amazonaws.com
The authenticity of host 'ec2-44-203-246-127.compute-1.amazonaws.com (44.203.246.127)' can't be established.
ED25519 key fingerprint is SHA256:CQU3TIN1zVgKmC7o2fj227mlzATLbOszRoiUcJbCaVw.
This key is not known by any other names
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'ec2-44-203-246-127.compute-1.amazonaws.com' (ED25519) to the list of known hosts.

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

https://aws.amazon.com/amazon-linux-2/
1 package(s) needed for security, out of 7 available
Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEEE 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
EEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR
```

Making Changes in consume.py file.

Users > rishabhjain > Desktop > consume.py

```python
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("ec2-44-203-246-127.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
```

Rishabh Jain || A20495530 || rjain35@hawk.iit.edu
Performing SCP for sending the files in cluster

```
→  Desktop scp -i /Users/rishabhjain/Desktop/new-key-pair-emr.pem /Users/rishabhjain/Desktop/consume.py hadoop@ec2-44-203-246-127.compute-1.a
mazonaws.com:/home/hadoop
consume.py
                100%  691     1.8KB/s   00:00
→  Desktop scp -i /Users/rishabhjain/Desktop/new-key-pair-emr.pem /Users/rishabhjain/Desktop/log4j.properties hadoop@ec2-44-203-246-127.compu
te-1.amazonaws.com:/home/hadoop
log4j.properties
                100% 3199    79.2KB/s   00:00
```

Terminal 1 Input

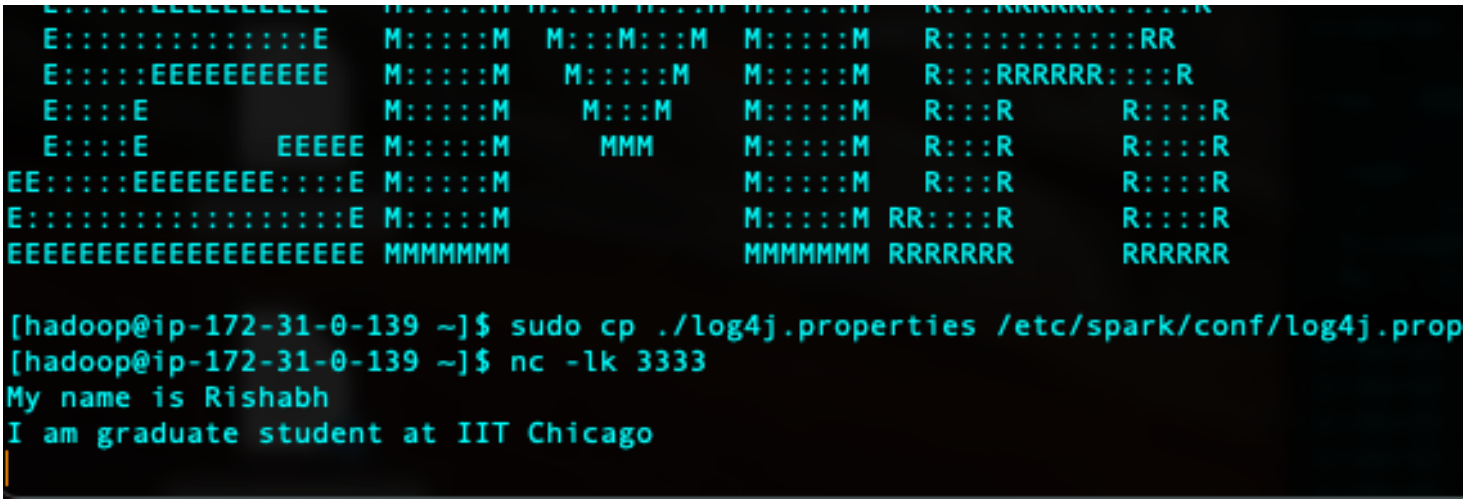
```
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   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
EEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-0-139 ~]$ sudo cp ./log4j.properties /etc/spark/conf/log4j.prop
[hadoop@ip-172-31-0-139 ~]$ nc -lk 3333
My name is Rishabh
```

Terminal 2 Output

```
22/04/03 22:52:30 INFO BlockManagerInfo: Added broadcast_25_piece0 in memory on i
B)
22/04/03 22:52:30 INFO TaskSetManager: Finished task 0.0 in stage 46.0 (TID 93) i
22/04/03 22:52:30 INFO YarnScheduler: Removed TaskSet 46.0, whose tasks have all
22/04/03 22:52:30 INFO DAGScheduler: ResultStage 46 (runJob at PythonRDD.scala:15
22/04/03 22:52:30 INFO DAGScheduler: Job 23 finished: runJob at PythonRDD.scala:1
-------------------------------------------
Time: 2022-04-03 22:52:30
-------------------------------------------
('name', 1)
('is', 1)
('Rishabh', 1)
('My', 1)
```

Rishabh Jain || A20495530 || rjain35@hawk.iit.edu

Terminal 1 Input



Terminal 2 Output