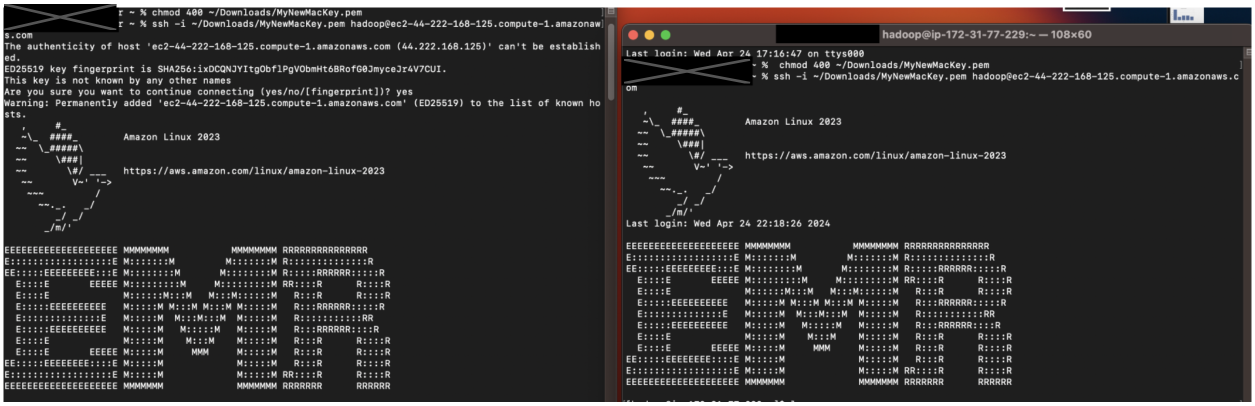
PART A:

STARTING LAB & CREATING CLUSTER:

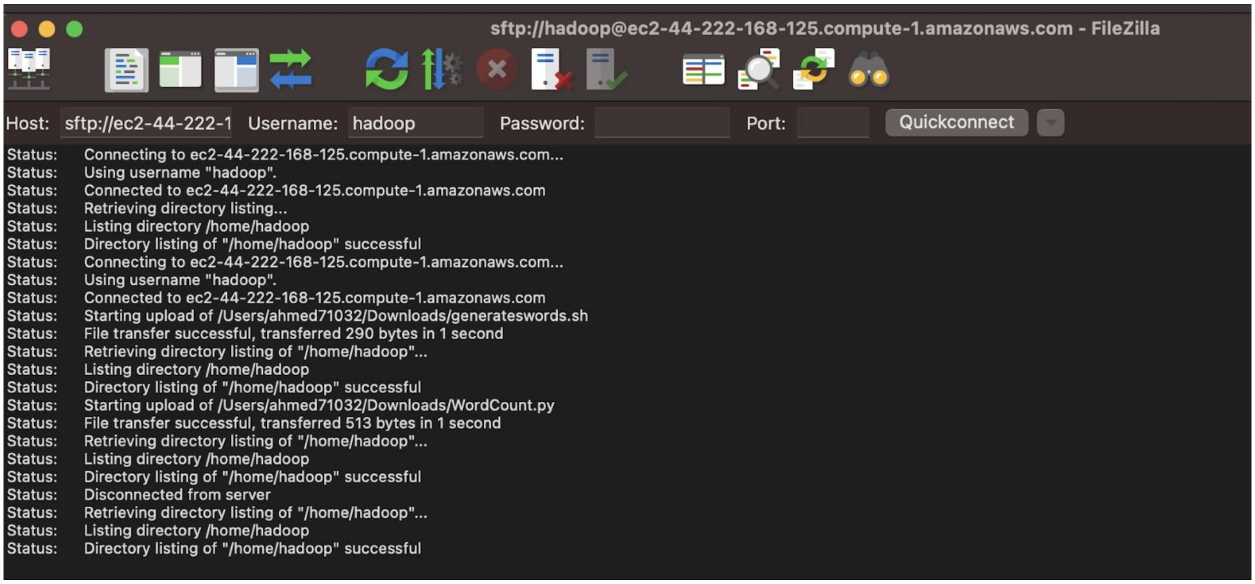


```
[redacted] ~ % chmod 400 ~/Downloads/MyNewMacKey.pem
 ~ % ssh -i ~/Downloads/MyNewMacKey.pem hadoop@ec2-44-222-168-125.compute-1.amazonaw
s.com
The authenticity of host 'ec2-44-222-168-125.compute-1.amazonaws.com (44.222.168.125)' can't be establish
ed.
ED25519 key fingerprint is SHA256:ixDCQNJYItgObflPgVObmHt6BRofG0JmyceJr4V7CUI.
This key is not known by any other names
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'ec2-44-222-168-125.compute-1.amazonaws.com' (ED25519) to the list of known ho
sts.
   ,     #_
   ~\_  ####_        Amazon Linux 2023
  ~~  \_#####\
  ~~     \###|
  ~~       \#/ ___   https://aws.amazon.com/linux/amazon-linux-2023
   ~~       V~' '->
    ~~~         /
      ~~._.   _/
         _/ _/
       _/m/'

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
```

[redacted] hadoop@ip-172-31-77-229:~ — 108×60

```
Last login: Wed Apr 24 17:16:47 on ttys000
[redacted] %  chmod 400 ~/Downloads/MyNewMacKey.pem
 % ssh -i ~/Downloads/MyNewMacKey.pem hadoop@ec2-44-222-168-125.compute-1.amazonaws.c
om

   ,     #_
   ~\_  ####_        Amazon Linux 2023
  ~~  \_#####\
  ~~     \###|
  ~~       \#/ ___   https://aws.amazon.com/linux/amazon-linux-2023
   ~~       V~' '->
    ~~~         /
      ~~._.   _/
         _/ _/
       _/m/'
Last login: Wed Apr 24 22:18:26 2024

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
```

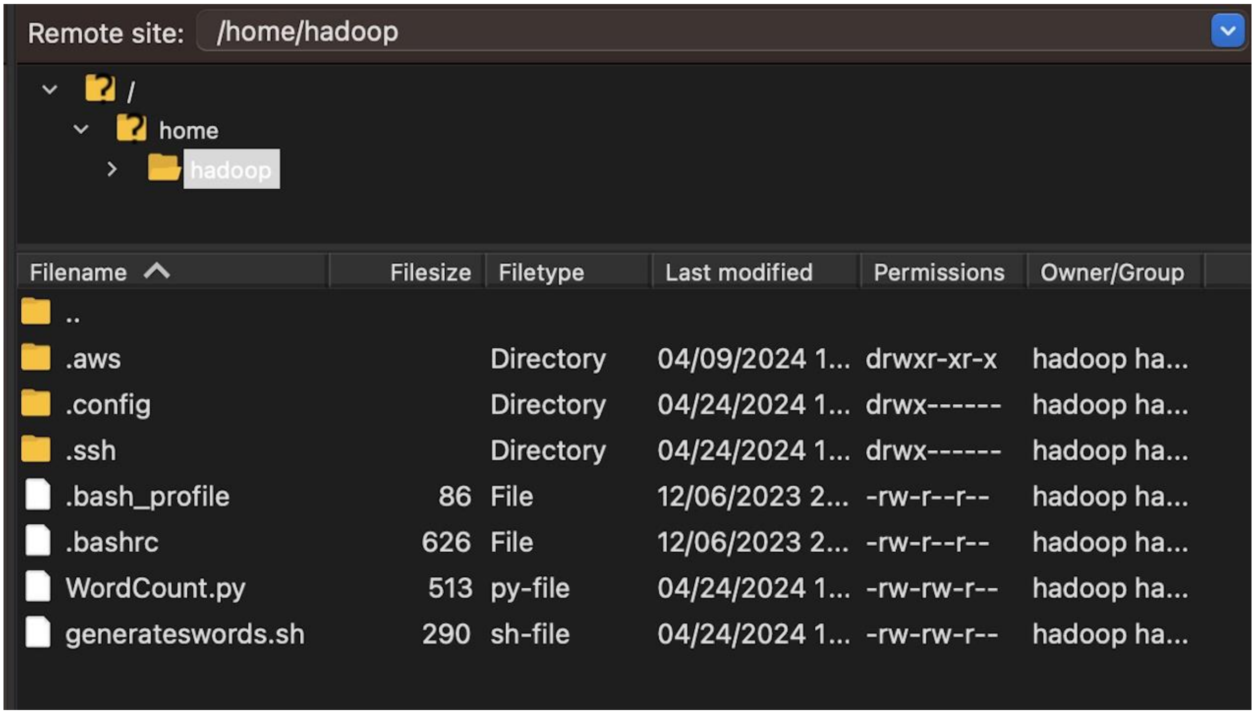
CONNECTING TO HADOOP:



sftp://hadoop@ec2-44-222-168-125.compute-1.amazonaws.com - FileZilla

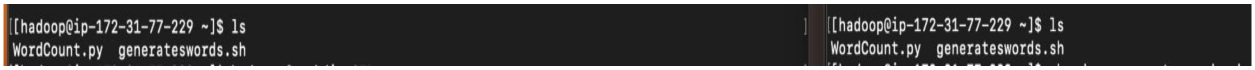
Host: sftp://ec2-44-222-1 Username: hadoop Password: Port:

```
Status:    Connecting to ec2-44-222-168-125.compute-1.amazonaws.com...
Status:    Using username "hadoop".
Status:    Connected to ec2-44-222-168-125.compute-1.amazonaws.com
Status:    Retrieving directory listing...
Status:    Listing directory /home/hadoop
Status:    Directory listing of "/home/hadoop" successful
Status:    Connecting to ec2-44-222-168-125.compute-1.amazonaws.com...
Status:    Using username "hadoop".
Status:    Connected to ec2-44-222-168-125.compute-1.amazonaws.com
Status:    Starting upload of /Users/ahmed71032/Downloads/generateswords.sh
Status:    File transfer successful, transferred 290 bytes in 1 second
Status:    Retrieving directory listing of "/home/hadoop"...
Status:    Listing directory /home/hadoop
Status:    Directory listing of "/home/hadoop" successful
Status:    Starting upload of /Users/ahmed71032/Downloads/WordCount.py
Status:    File transfer successful, transferred 513 bytes in 1 second
Status:    Retrieving directory listing of "/home/hadoop"...
Status:    Listing directory /home/hadoop
Status:    Directory listing of "/home/hadoop" successful
Status:    Disconnected from server
Status:    Retrieving directory listing of "/home/hadoop"...
Status:    Listing directory /home/hadoop
Status:    Directory listing of "/home/hadoop" successful
```



Remote site: /home/hadoop

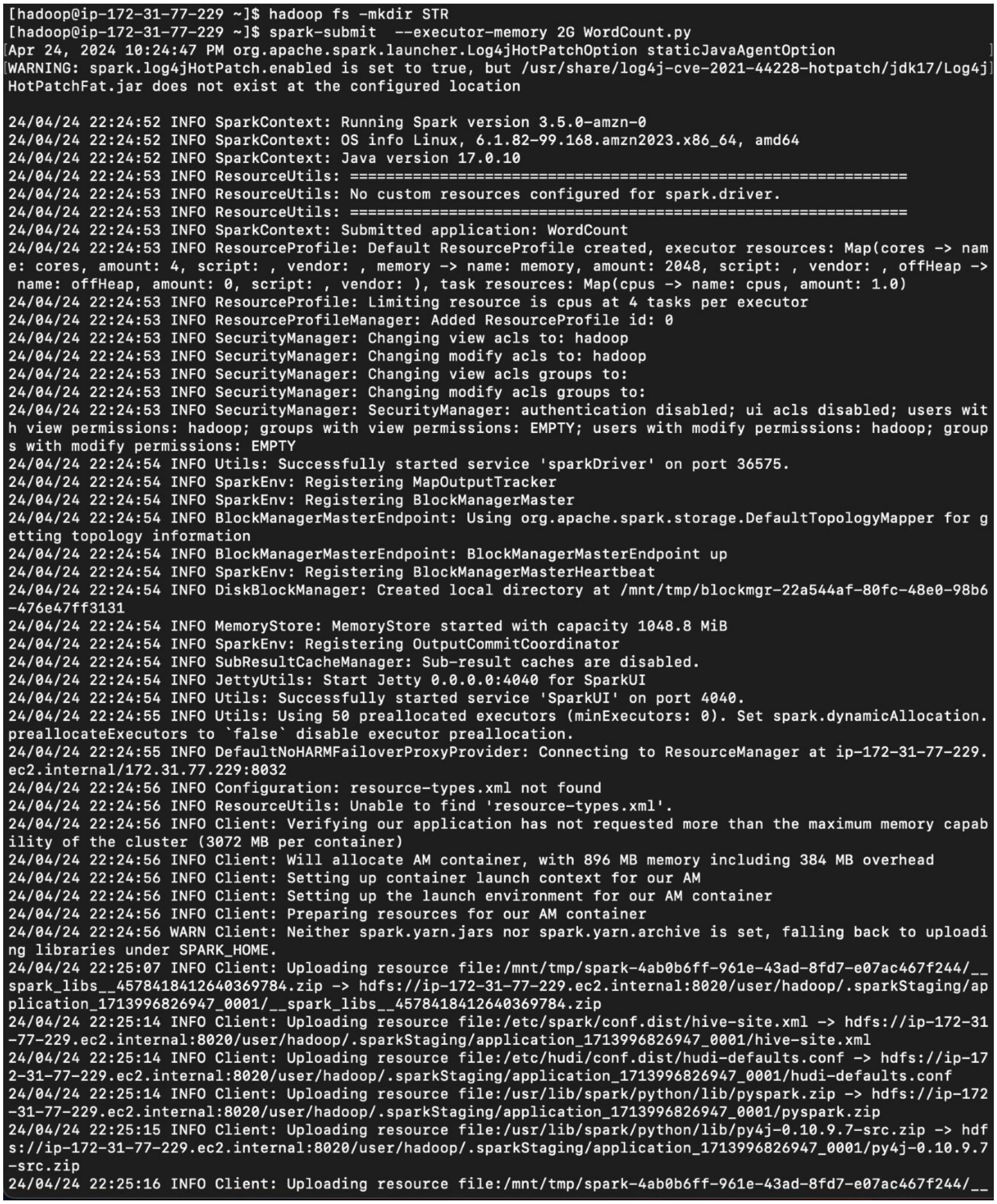
Filename	Filesize	Filetype	Last modified	Permissions	Owner/Group
..					
.aws		Directory	04/09/2024 1...	drwxr-xr-x	hadoop ha...
.config		Directory	04/24/2024 1...	drwx------	hadoop ha...
.ssh		Directory	04/24/2024 1...	drwx------	hadoop ha...
.bash_profile	86	File	12/06/2023 2...	-rw-r--r--	hadoop ha...
.bashrc	626	File	12/06/2023 2...	-rw-r--r--	hadoop ha...
WordCount.py	513	py-file	04/24/2024 1...	-rw-rw-r--	hadoop ha...
generateswords.sh	290	sh-file	04/24/2024 1...	-rw-rw-r--	hadoop ha...



```
[[hadoop@ip-172-31-77-229 ~]$ ls
 WordCount.py  generateswords.sh
```

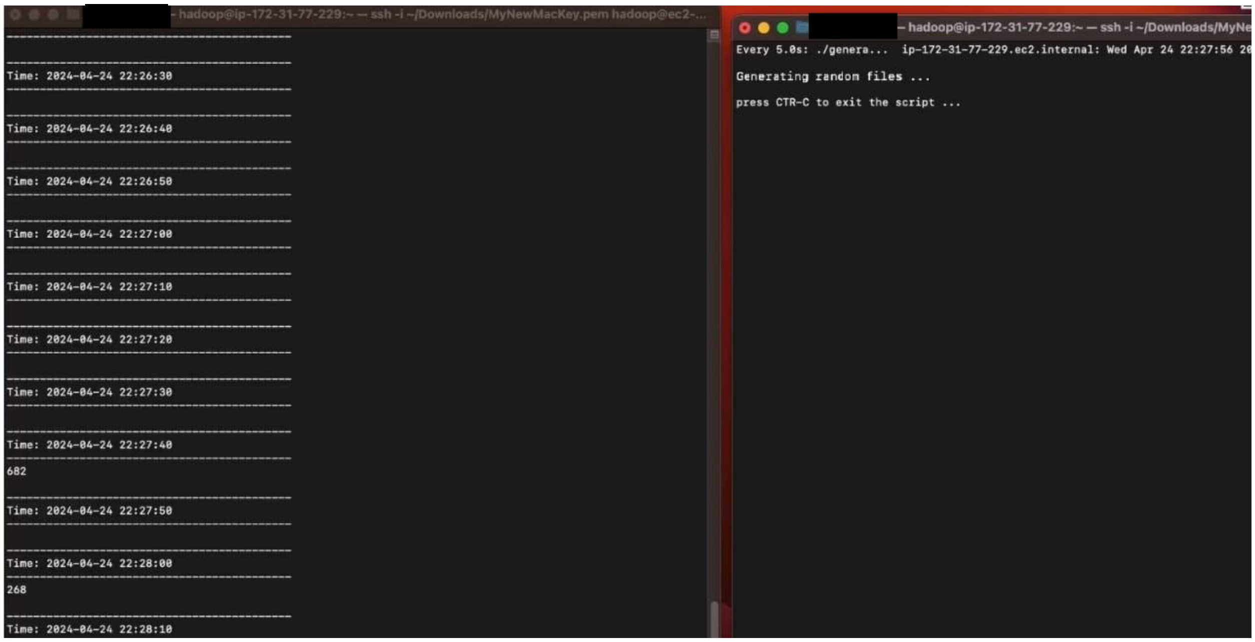
```
[[hadoop@ip-172-31-77-229 ~]$ ls
 WordCount.py  generateswords.sh
```

CREATING DIRECTORY AND LAUNCHING SPARK:



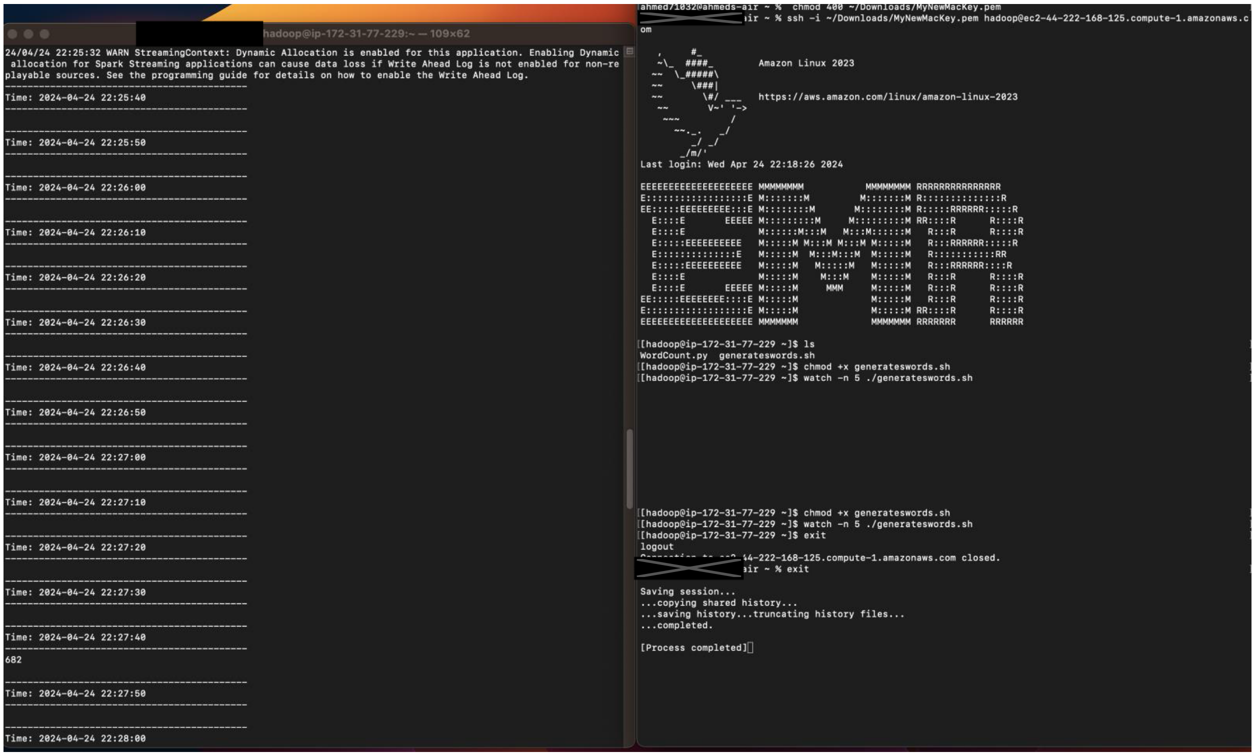
```
[hadoop@ip-172-31-77-229 ~]$ hadoop fs -mkdir STR
[hadoop@ip-172-31-77-229 ~]$ spark-submit  --executor-memory 2G WordCount.py
[Apr 24, 2024 10:24:47 PM org.apache.spark.launcher.Log4jHotPatchOption staticJavaAgentOption         ]
[WARNING: spark.log4jHotPatch.enabled is set to true, but /usr/share/log4j-cve-2021-44228-hotpatch/jdk17/Log4j]
HotPatchFat.jar does not exist at the configured location

24/04/24 22:24:52 INFO SparkContext: Running Spark version 3.5.0-amzn-0
24/04/24 22:24:52 INFO SparkContext: OS info Linux, 6.1.82-99.168.amzn2023.x86_64, amd64
24/04/24 22:24:52 INFO SparkContext: Java version 17.0.10
24/04/24 22:24:53 INFO ResourceUtils: ==============================================================
24/04/24 22:24:53 INFO ResourceUtils: No custom resources configured for spark.driver.
24/04/24 22:24:53 INFO ResourceUtils: ==============================================================
24/04/24 22:24:53 INFO SparkContext: Submitted application: WordCount
24/04/24 22:24:53 INFO ResourceProfile: Default ResourceProfile created, executor resources: Map(cores -> nam
e: cores, amount: 4, script: , vendor: , memory -> name: memory, amount: 2048, script: , vendor: , offHeap ->
 name: offHeap, amount: 0, script: , vendor: ), task resources: Map(cpus -> name: cpus, amount: 1.0)
24/04/24 22:24:53 INFO ResourceProfile: Limiting resource is cpus at 4 tasks per executor
24/04/24 22:24:53 INFO ResourceProfileManager: Added ResourceProfile id: 0
24/04/24 22:24:53 INFO SecurityManager: Changing view acls to: hadoop
24/04/24 22:24:53 INFO SecurityManager: Changing modify acls to: hadoop
24/04/24 22:24:53 INFO SecurityManager: Changing view acls groups to:
24/04/24 22:24:53 INFO SecurityManager: Changing modify acls groups to:
24/04/24 22:24:53 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users wit
h view permissions: hadoop; groups with view permissions: EMPTY; users with modify permissions: hadoop; group
s with modify permissions: EMPTY
24/04/24 22:24:54 INFO Utils: Successfully started service 'sparkDriver' on port 36575.
24/04/24 22:24:54 INFO SparkEnv: Registering MapOutputTracker
24/04/24 22:24:54 INFO SparkEnv: Registering BlockManagerMaster
24/04/24 22:24:54 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for g
etting topology information
24/04/24 22:24:54 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
24/04/24 22:24:54 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
24/04/24 22:24:54 INFO DiskBlockManager: Created local directory at /mnt/tmp/blockmgr-22a544af-80fc-48e0-98b6
-476e47ff3131
24/04/24 22:24:54 INFO MemoryStore: MemoryStore started with capacity 1048.8 MiB
24/04/24 22:24:54 INFO SparkEnv: Registering OutputCommitCoordinator
24/04/24 22:24:54 INFO SubResultCacheManager: Sub-result caches are disabled.
24/04/24 22:24:54 INFO JettyUtils: Start Jetty 0.0.0.0:4040 for SparkUI
24/04/24 22:24:54 INFO Utils: Successfully started service 'SparkUI' on port 4040.
24/04/24 22:24:55 INFO Utils: Using 50 preallocated executors (minExecutors: 0). Set spark.dynamicAllocation.
preallocateExecutors to `false` disable executor preallocation.
24/04/24 22:24:55 INFO DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at ip-172-31-77-229.
ec2.internal/172.31.77.229:8032
24/04/24 22:24:56 INFO Configuration: resource-types.xml not found
24/04/24 22:24:56 INFO ResourceUtils: Unable to find 'resource-types.xml'.
24/04/24 22:24:56 INFO Client: Verifying our application has not requested more than the maximum memory capab
ility of the cluster (3072 MB per container)
24/04/24 22:24:56 INFO Client: Will allocate AM container, with 896 MB memory including 384 MB overhead
24/04/24 22:24:56 INFO Client: Setting up container launch context for our AM
24/04/24 22:24:56 INFO Client: Setting up the launch environment for our AM container
24/04/24 22:24:56 INFO Client: Preparing resources for our AM container
24/04/24 22:24:56 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploadi
ng libraries under SPARK_HOME.
24/04/24 22:25:07 INFO Client: Uploading resource file:/mnt/tmp/spark-4ab0b6ff-961e-43ad-8fd7-e07ac467f244/__
spark_libs__4578418412640369784.zip -> hdfs://ip-172-31-77-229.ec2.internal:8020/user/hadoop/.sparkStaging/ap
plication_1713996826947_0001/__spark_libs__4578418412640369784.zip
24/04/24 22:25:14 INFO Client: Uploading resource file:/etc/spark/conf.dist/hive-site.xml -> hdfs://ip-172-31
-77-229.ec2.internal:8020/user/hadoop/.sparkStaging/application_1713996826947_0001/hive-site.xml
24/04/24 22:25:14 INFO Client: Uploading resource file:/etc/hudi/conf.dist/hudi-defaults.conf -> hdfs://ip-17
2-31-77-229.ec2.internal:8020/user/hadoop/.sparkStaging/application_1713996826947_0001/hudi-defaults.conf
24/04/24 22:25:14 INFO Client: Uploading resource file:/usr/lib/spark/python/lib/pyspark.zip -> hdfs://ip-172
-31-77-229.ec2.internal:8020/user/hadoop/.sparkStaging/application_1713996826947_0001/pyspark.zip
24/04/24 22:25:15 INFO Client: Uploading resource file:/usr/lib/spark/python/lib/py4j-0.10.9.7-src.zip -> hdf
s://ip-172-31-77-229.ec2.internal:8020/user/hadoop/.sparkStaging/application_1713996826947_0001/py4j-0.10.9.7
-src.zip
24/04/24 22:25:16 INFO Client: Uploading resource file:/mnt/tmp/spark-4ab0b6ff-961e-43ad-8fd7-e07ac467f244/__
```

GENERATING RANDOM NUMBERS:



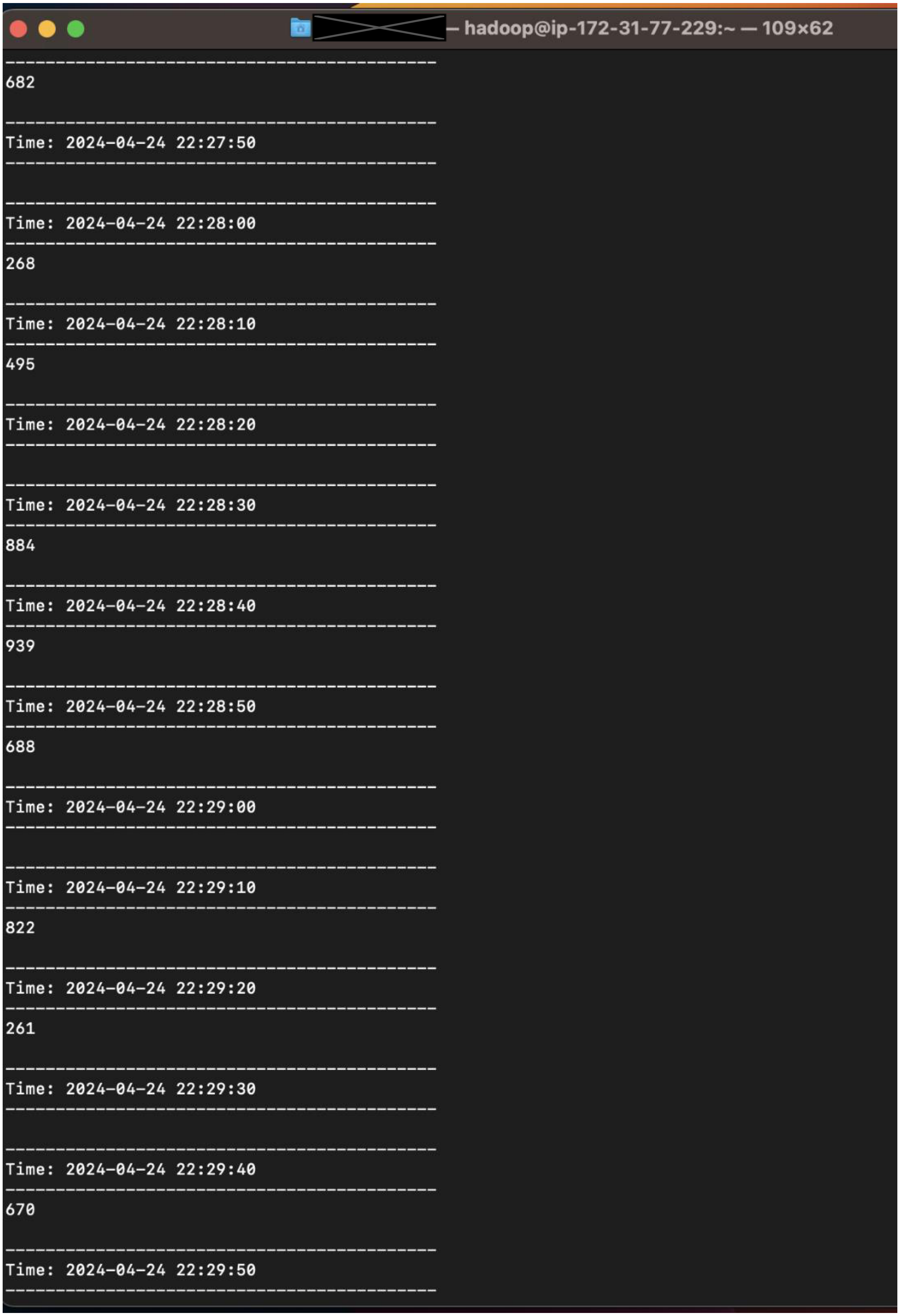
```
[redacted] - hadoop@ip-172-31-77-229:~ — ssh -i ~/Downloads/MyNewMacKey.pem hadoop@ec2-...
-------------------------------------------

-------------------------------------------
Time: 2024-04-24 22:26:30
-------------------------------------------

-------------------------------------------
Time: 2024-04-24 22:26:40
-------------------------------------------

-------------------------------------------
Time: 2024-04-24 22:26:50
-------------------------------------------

-------------------------------------------
Time: 2024-04-24 22:27:00
-------------------------------------------

-------------------------------------------
Time: 2024-04-24 22:27:10
-------------------------------------------

-------------------------------------------
Time: 2024-04-24 22:27:20
-------------------------------------------

-------------------------------------------
Time: 2024-04-24 22:27:30
-------------------------------------------

-------------------------------------------
Time: 2024-04-24 22:27:40
-------------------------------------------
682

-------------------------------------------
Time: 2024-04-24 22:27:50
-------------------------------------------

-------------------------------------------
Time: 2024-04-24 22:28:00
-------------------------------------------
268

-------------------------------------------
Time: 2024-04-24 22:28:10
```

```
[redacted] - hadoop@ip-172-31-77-229:~ — ssh -i ~/Downloads/MyNe
Every 5.0s: ./genera...  ip-172-31-77-229.ec2.internal: Wed Apr 24 22:27:56 20

Generating random files ...

press CTR-C to exit the script ...
```



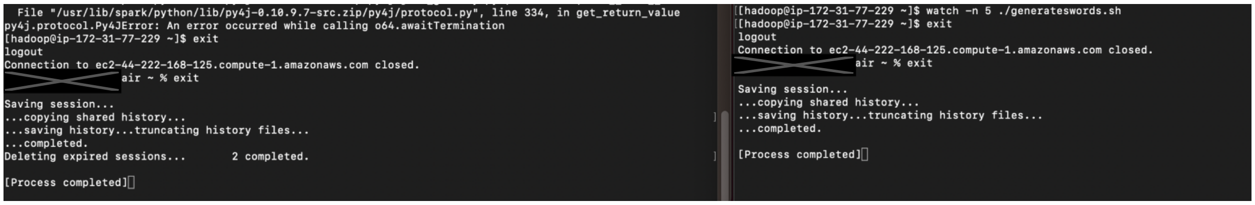
```
[redacted] hadoop@ip-172-31-77-229:~ — 109×62
24/04/24 22:25:32 WARN StreamingContext: Dynamic Allocation is enabled for this application. Enabling Dynamic
 allocation for Spark Streaming applications can cause data loss if Write Ahead Log is not enabled for non-re
playable sources. See the programming guide for details on how to enable the Write Ahead Log.
-------------------------------------------
Time: 2024-04-24 22:25:40
-------------------------------------------

-------------------------------------------
Time: 2024-04-24 22:25:50
-------------------------------------------

-------------------------------------------
Time: 2024-04-24 22:26:00
-------------------------------------------

-------------------------------------------
Time: 2024-04-24 22:26:10
-------------------------------------------

-------------------------------------------
Time: 2024-04-24 22:26:20
-------------------------------------------

-------------------------------------------
Time: 2024-04-24 22:26:30
-------------------------------------------

-------------------------------------------
Time: 2024-04-24 22:26:40
-------------------------------------------

-------------------------------------------
Time: 2024-04-24 22:26:50
-------------------------------------------

-------------------------------------------
Time: 2024-04-24 22:27:00
-------------------------------------------

-------------------------------------------
Time: 2024-04-24 22:27:10
-------------------------------------------

-------------------------------------------
Time: 2024-04-24 22:27:20
-------------------------------------------

-------------------------------------------
Time: 2024-04-24 22:27:30
-------------------------------------------

-------------------------------------------
Time: 2024-04-24 22:27:40
-------------------------------------------
682

-------------------------------------------
Time: 2024-04-24 22:27:50
-------------------------------------------

-------------------------------------------
Time: 2024-04-24 22:28:00
```

```
ahmed71032@ahmeds-air ~ %  chmod 400 ~/Downloads/MyNewMacKey.pem
[redacted]air ~ % ssh -i ~/Downloads/MyNewMacKey.pem hadoop@ec2-44-222-168-125.compute-1.amazonaws.c
om

   ,     #_
   ~\_  ####_        Amazon Linux 2023
  ~~  \_#####\
  ~~     \###|
  ~~       \#/ ___   https://aws.amazon.com/linux/amazon-linux-2023
   ~~       V~' '->
    ~~~         /
      ~~._.   _/
         _/ _/
       _/m/'
Last login: Wed Apr 24 22:18:26 2024

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-172-31-77-229 ~]$ ls
WordCount.py  generateswords.sh
[hadoop@ip-172-31-77-229 ~]$ chmod +x generateswords.sh
[hadoop@ip-172-31-77-229 ~]$ watch -n 5 ./generateswords.sh








[hadoop@ip-172-31-77-229 ~]$ chmod +x generateswords.sh
[hadoop@ip-172-31-77-229 ~]$ watch -n 5 ./generateswords.sh
[hadoop@ip-172-31-77-229 ~]$ exit
logout
[redacted]44-222-168-125.compute-1.amazonaws.com closed.
air ~ % exit

Saving session...
...copying shared history...
...saving history...truncating history files...
...completed.

[Process completed]
```



```
-------------------------------------------
682

-------------------------------------------
Time: 2024-04-24 22:27:50
-------------------------------------------

-------------------------------------------
Time: 2024-04-24 22:28:00
-------------------------------------------
268

-------------------------------------------
Time: 2024-04-24 22:28:10
-------------------------------------------
495

-------------------------------------------
Time: 2024-04-24 22:28:20
-------------------------------------------

-------------------------------------------
Time: 2024-04-24 22:28:30
-------------------------------------------
884

-------------------------------------------
Time: 2024-04-24 22:28:40
-------------------------------------------
939

-------------------------------------------
Time: 2024-04-24 22:28:50
-------------------------------------------
688

-------------------------------------------
Time: 2024-04-24 22:29:00
-------------------------------------------

-------------------------------------------
Time: 2024-04-24 22:29:10
-------------------------------------------
822

-------------------------------------------
Time: 2024-04-24 22:29:20
-------------------------------------------
261

-------------------------------------------
Time: 2024-04-24 22:29:30
-------------------------------------------

-------------------------------------------
Time: 2024-04-24 22:29:40
-------------------------------------------
670

-------------------------------------------
Time: 2024-04-24 22:29:50
-------------------------------------------
```

EXIT FROM THE MACHINE:



```
  File "/usr/lib/spark/python/lib/py4j-0.10.9.7-src.zip/py4j/protocol.py", line 334, in get_return_value
py4j.protocol.Py4JError: An error occurred while calling o64.awaitTermination
[hadoop@ip-172-31-77-229 ~]$ exit
logout
Connection to ec2-44-222-168-125.compute-1.amazonaws.com closed.
[redacted]air ~ % exit

Saving session...
...copying shared history...
...saving history...truncating history files...
...completed.
Deleting expired sessions...       2 completed.

[Process completed]
```

```
[hadoop@ip-172-31-77-229 ~]$ watch -n 5 ./generateswords.sh
[hadoop@ip-172-31-77-229 ~]$ exit
logout
Connection to ec2-44-222-168-125.compute-1.amazonaws.com closed.
[redacted]air ~ % exit

Saving session...
...copying shared history...
...saving history...truncating history files...
...completed.

[Process completed]
```