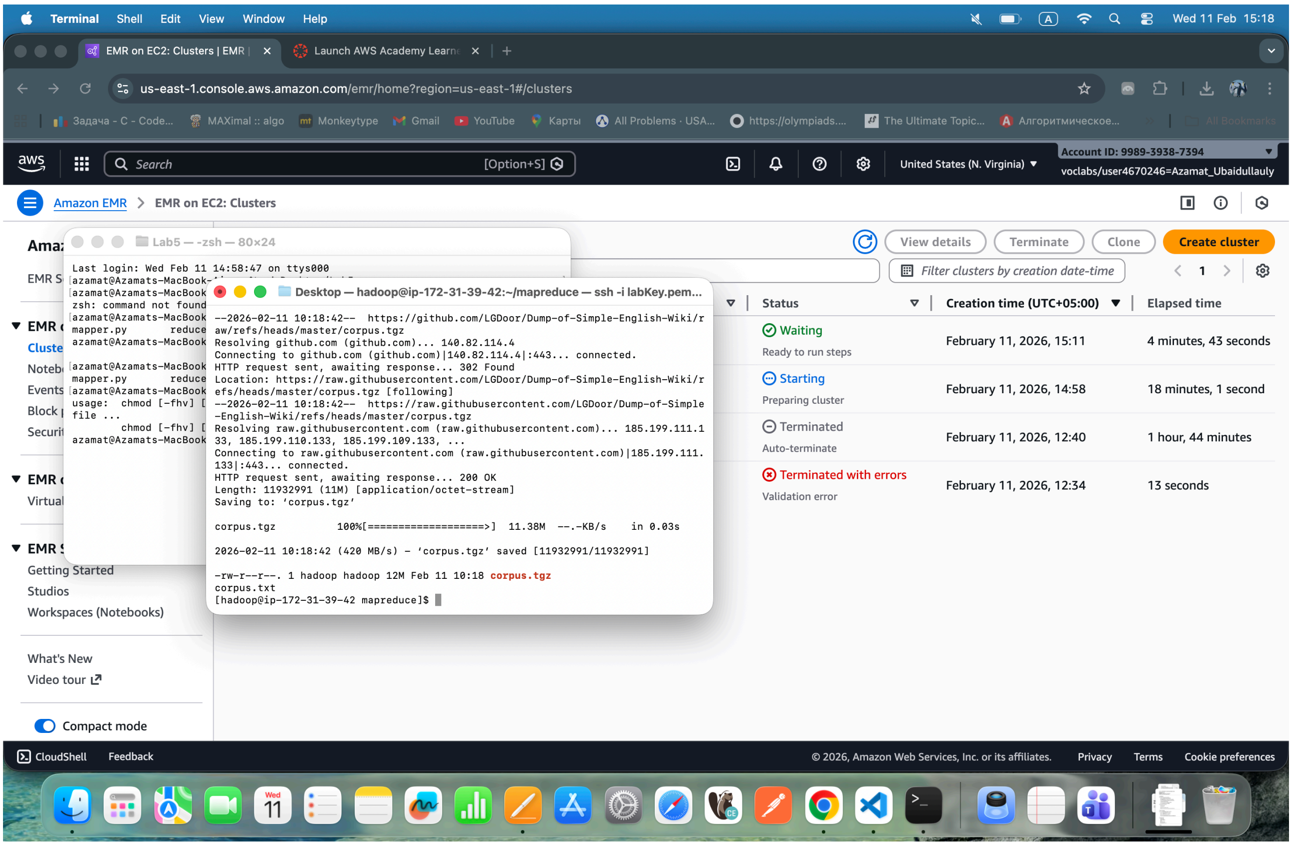
Lab 5 Report
Student: Ubaidullauly Azamat
Group: IT-2305

I created a cluster and installed a dataset via this command:

```
# on master node
cd ~
mkdir -p mapreduce && cd mapreduce

wget https://github.com/LGDoor/Dump-of-Simple-English-Wiki/raw/refs/heads/master/corpus.tgz
ls -lh corpus.tgz
tar -xvzf corpus.tgz
```



Lab5 — -zsh — 80×24

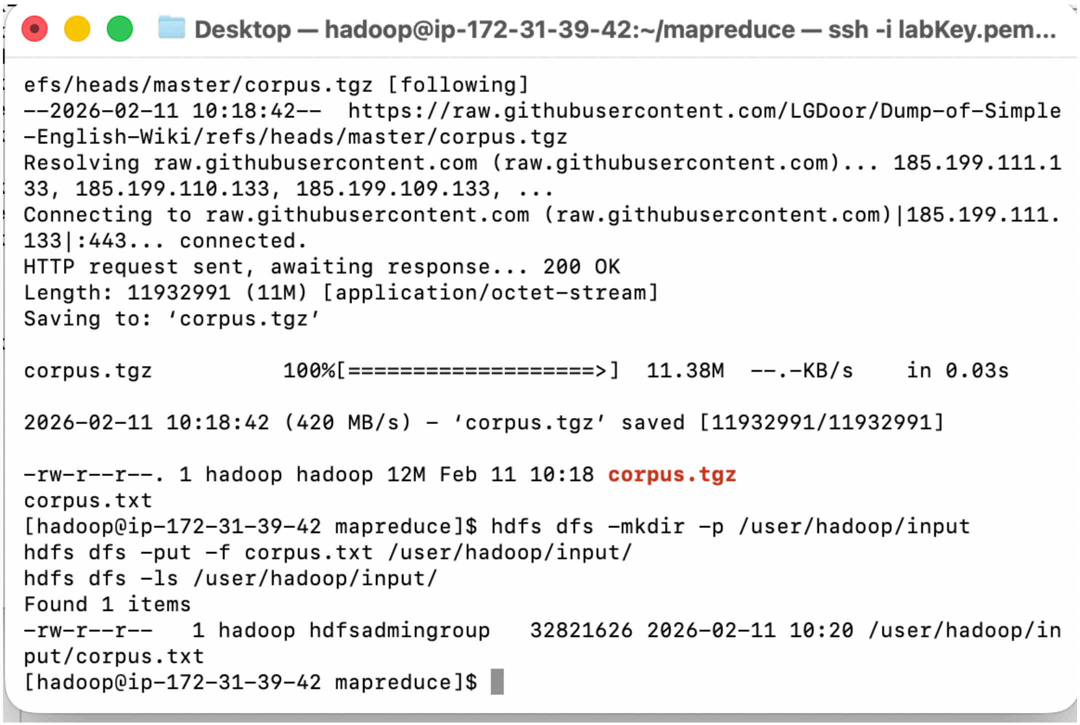
```
Last login: Wed Feb 11 14:58:47 on ttys000
azamat@Azamats-MacBook-
azamat@Azamats-MacBook
zsh: command not found
azamat@Azamats-MacBook
mapper.py        reduce
azamat@Azamats-MacBook

azamat@Azamats-MacBook
mapper.py        reduce
azamat@Azamats-MacBook
usage:  chmod [-fhv] [
file ...
        chmod [-fhv] [
azamat@Azamats-MacBook
```

Desktop — hadoop@ip-172-31-39-42:~/mapreduce — ssh -i labKey.pem...

```
--2026-02-11 10:18:42--  https://github.com/LGDoor/Dump-of-Simple-English-Wiki/r
aw/refs/heads/master/corpus.tgz
Resolving github.com (github.com)... 140.82.114.4
Connecting to github.com (github.com)|140.82.114.4|:443... connected.
HTTP request sent, awaiting response... 302 Found
Location: https://raw.githubusercontent.com/LGDoor/Dump-of-Simple-English-Wiki/r
efs/heads/master/corpus.tgz [following]
--2026-02-11 10:18:42--  https://raw.githubusercontent.com/LGDoor/Dump-of-Simple
-English-Wiki/refs/heads/master/corpus.tgz
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.111.1
33, 185.199.110.133, 185.199.109.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.111.
133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 11932991 (11M) [application/octet-stream]
Saving to: ‘corpus.tgz’

corpus.tgz          100%[===================>]  11.38M  --.-KB/s    in 0.03s

2026-02-11 10:18:42 (420 MB/s) - ‘corpus.tgz’ saved [11932991/11932991]

-rw-r--r--. 1 hadoop hadoop 12M Feb 11 10:18 corpus.tgz
corpus.txt
[hadoop@ip-172-31-39-42 mapreduce]$
```

Status	Creation time (UTC+05:00)	Elapsed time
Waiting Ready to run steps	February 11, 2026, 15:11	4 minutes, 43 seconds
Starting Preparing cluster	February 11, 2026, 14:58	18 minutes, 1 second
Terminated Auto-terminate	February 11, 2026, 12:40	1 hour, 44 minutes
Terminated with errors Validation error	February 11, 2026, 12:34	13 seconds

Then we put it into HDFS using the commands below:

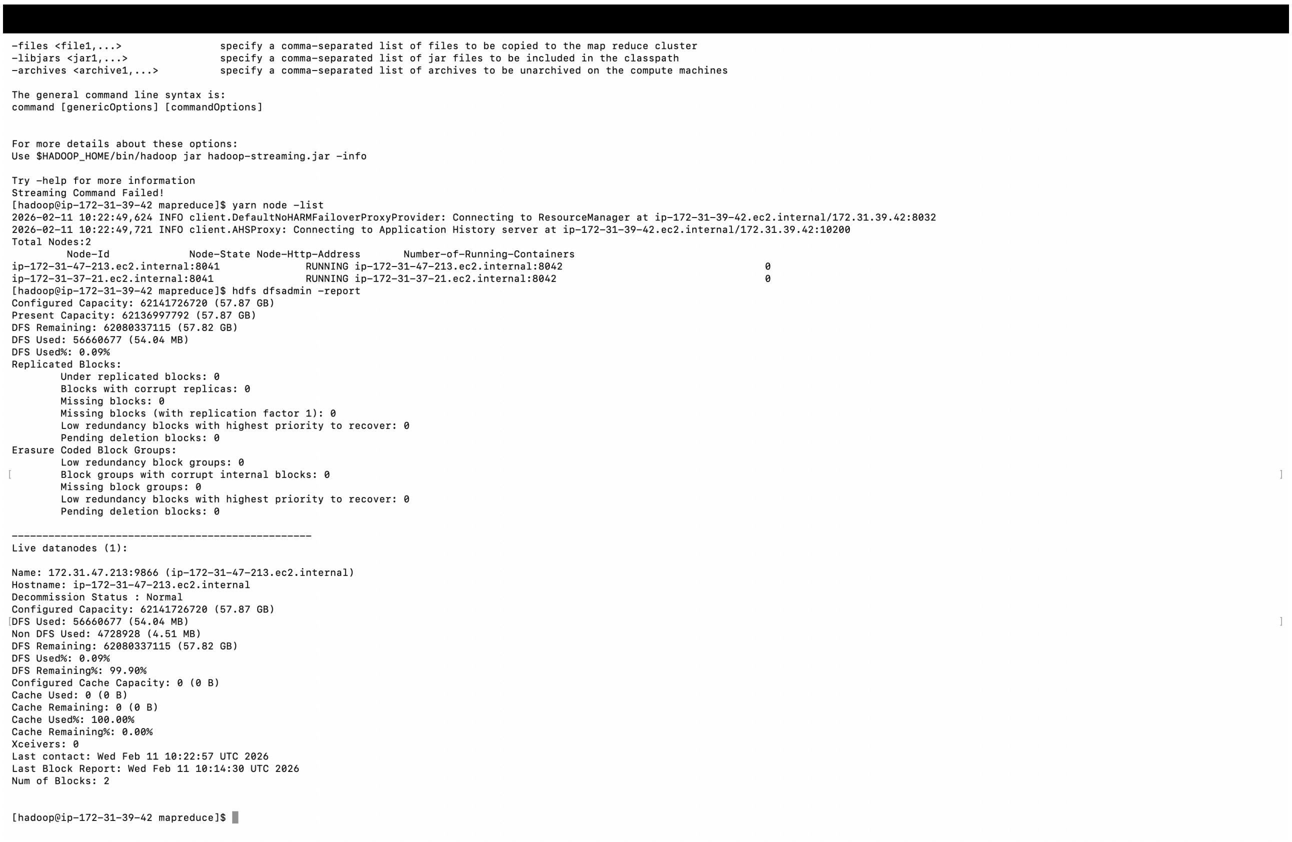


Desktop — hadoop@ip-172-31-39-42:~/mapreduce — ssh -i labKey.pem...

```
efs/heads/master/corpus.tgz [following]
--2026-02-11 10:18:42--  https://raw.githubusercontent.com/LGDoor/Dump-of-Simple
-English-Wiki/refs/heads/master/corpus.tgz
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.111.1
33, 185.199.110.133, 185.199.109.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.111.
133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 11932991 (11M) [application/octet-stream]
Saving to: ‘corpus.tgz’

corpus.tgz          100%[===================>]  11.38M  --.-KB/s    in 0.03s

2026-02-11 10:18:42 (420 MB/s) - ‘corpus.tgz’ saved [11932991/11932991]

-rw-r--r--. 1 hadoop hadoop 12M Feb 11 10:18 corpus.tgz
corpus.txt
[hadoop@ip-172-31-39-42 mapreduce]$ hdfs dfs -mkdir -p /user/hadoop/input
hdfs dfs -put -f corpus.txt /user/hadoop/input/
hdfs dfs -ls /user/hadoop/input/
Found 1 items
-rw-r--r--   1 hadoop hdfsadmingroup   32821626 2026-02-11 10:20 /user/hadoop/in
put/corpus.txt
[hadoop@ip-172-31-39-42 mapreduce]$
```

```
hdfs dfs -mkdir -p /user/hadoop/input
hdfs dfs -put -f corpus.txt /user/hadoop/input/
hdfs dfs -ls /user/hadoop/input/
```

Check cluster if it is running:



[redacted]

```
-files <file1,...>                specify a comma-separated list of files to be copied to the map reduce cluster
-libjars <jar1,...>               specify a comma-separated list of jar files to be included in the classpath
-archives <archive1,...>          specify a comma-separated list of archives to be unarchived on the compute machines

The general command line syntax is:
command [genericOptions] [commandOptions]


For more details about these options:
Use $HADOOP_HOME/bin/hadoop jar hadoop-streaming.jar -info

Try -help for more information
Streaming Command Failed!
[hadoop@ip-172-31-39-42 mapreduce]$ yarn node -list
2026-02-11 10:22:49,624 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at ip-172-31-39-42.ec2.internal/172.31.39.42:8032
2026-02-11 10:22:49,721 INFO client.AHSProxy: Connecting to Application History server at ip-172-31-39-42.ec2.internal/172.31.39.42:10200
Total Nodes:2
         Node-Id             Node-State Node-Http-Address       Number-of-Running-Containers
ip-172-31-47-213.ec2.internal:8041              RUNNING ip-172-31-47-213.ec2.internal:8042                                0
ip-172-31-37-21.ec2.internal:8041               RUNNING ip-172-31-37-21.ec2.internal:8042                                 0
[hadoop@ip-172-31-39-42 mapreduce]$ hdfs dfsadmin -report
Configured Capacity: 62141726720 (57.87 GB)
Present Capacity: 62136997792 (57.87 GB)
DFS Remaining: 62080337115 (57.82 GB)
DFS Used: 56660677 (54.04 MB)
DFS Used%: 0.09%
Replicated Blocks:
        Under replicated blocks: 0
        Blocks with corrupt replicas: 0
        Missing blocks: 0
        Missing blocks (with replication factor 1): 0
        Low redundancy blocks with highest priority to recover: 0
        Pending deletion blocks: 0
Erasure Coded Block Groups:
        Low redundancy block groups: 0
        Block groups with corrupt internal blocks: 0
        Missing block groups: 0
        Low redundancy blocks with highest priority to recover: 0
        Pending deletion blocks: 0

-------------------------------------------------
Live datanodes (1):

Name: 172.31.47.213:9866 (ip-172-31-47-213.ec2.internal)
Hostname: ip-172-31-47-213.ec2.internal
Decommission Status : Normal
Configured Capacity: 62141726720 (57.87 GB)
DFS Used: 56660677 (54.04 MB)
Non DFS Used: 4728928 (4.51 MB)
DFS Remaining: 62080337115 (57.82 GB)
DFS Used%: 0.09%
DFS Remaining%: 99.90%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 0
Last contact: Wed Feb 11 10:22:57 UTC 2026
Last Block Report: Wed Feb 11 10:14:30 UTC 2026
Num of Blocks: 2


[hadoop@ip-172-31-39-42 mapreduce]$
```