AWS EMR:

EMR cluster is used to setup both Data Engineering and BI or adhoc Querying.

- I created the new EC2 key pair.
- Set up the EMR cluster with spark.

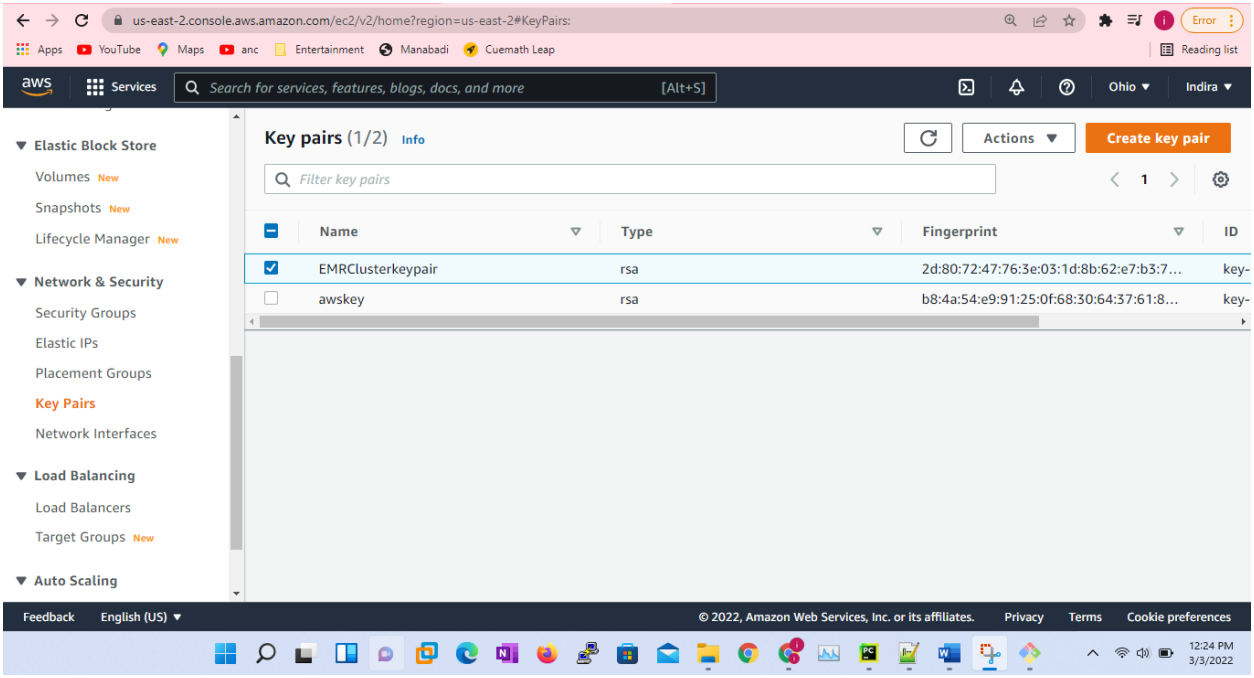


Figure 1Created new Keypair

Using EMR cluster ipadress started the terminal using SSH

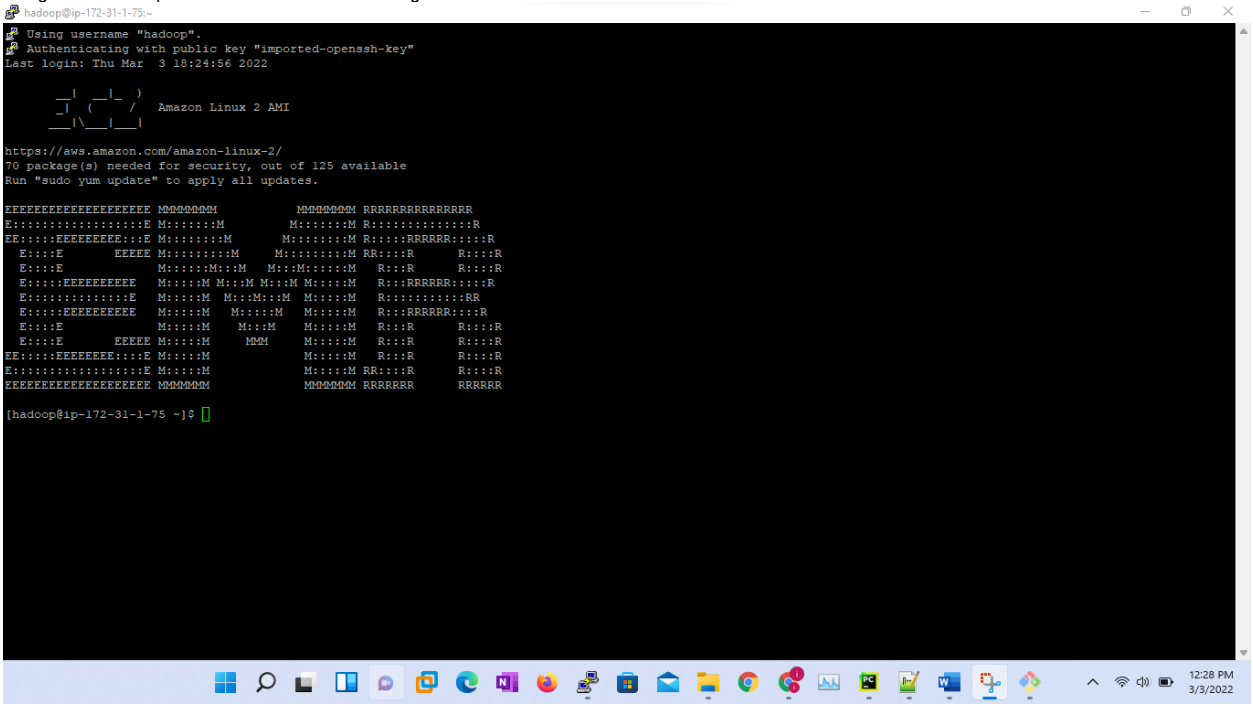
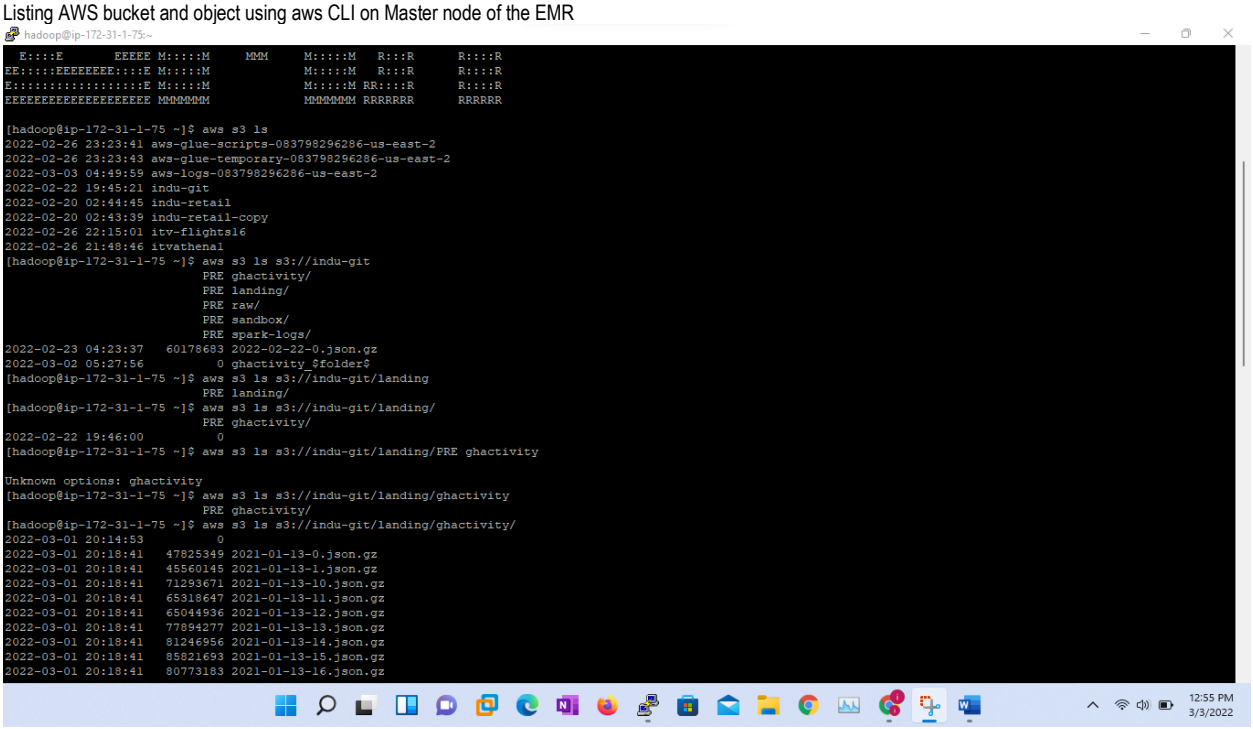
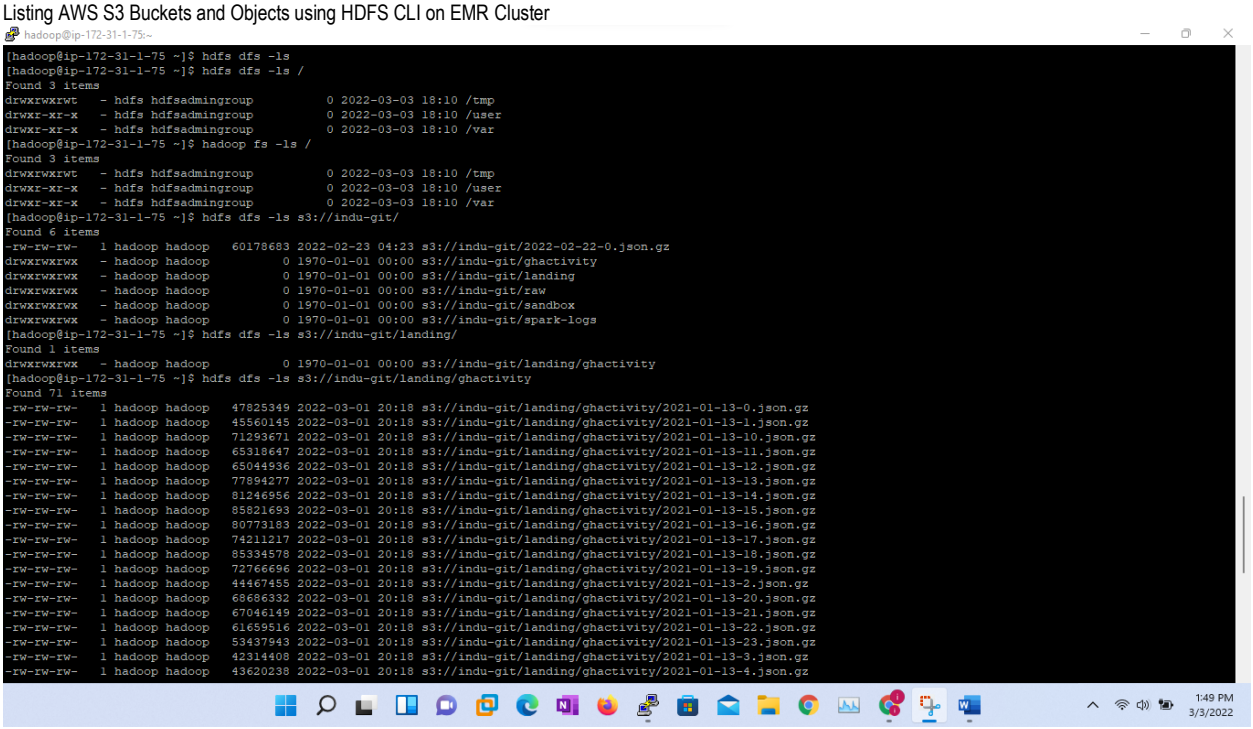


Figure 2Connected EMR master node using SSH

Listing AWS bucket and object using aws CLI on Master node of the EMR

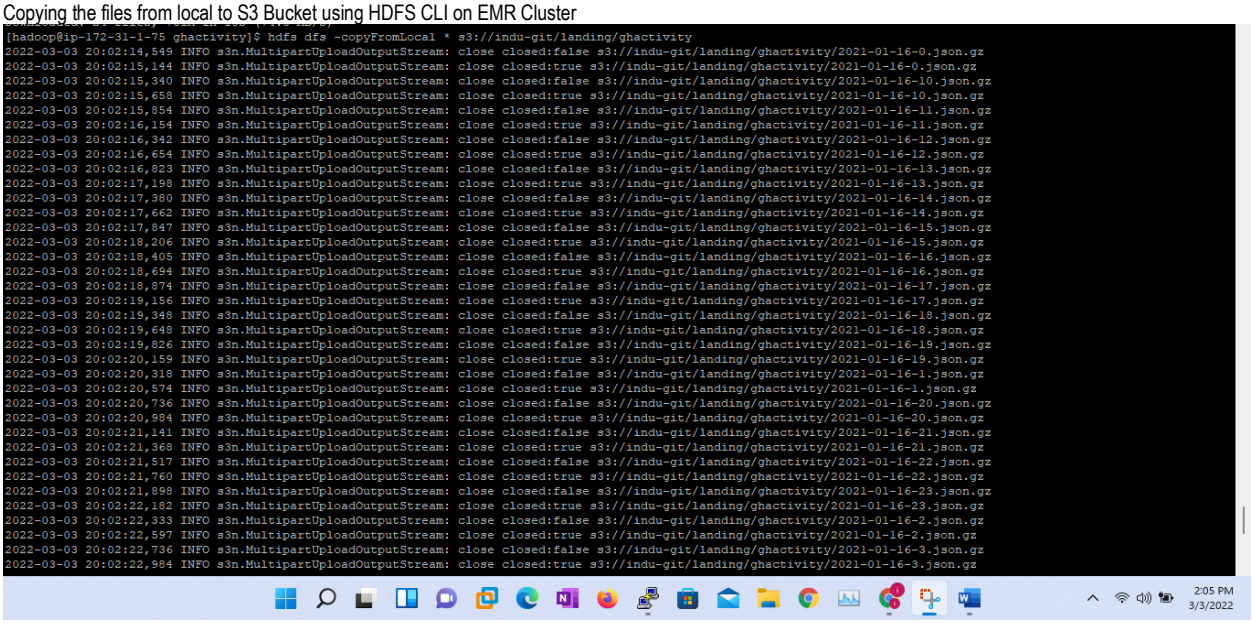


Listing AWS S3 Buckets and Objects using HDFS CLI on EMR Cluster

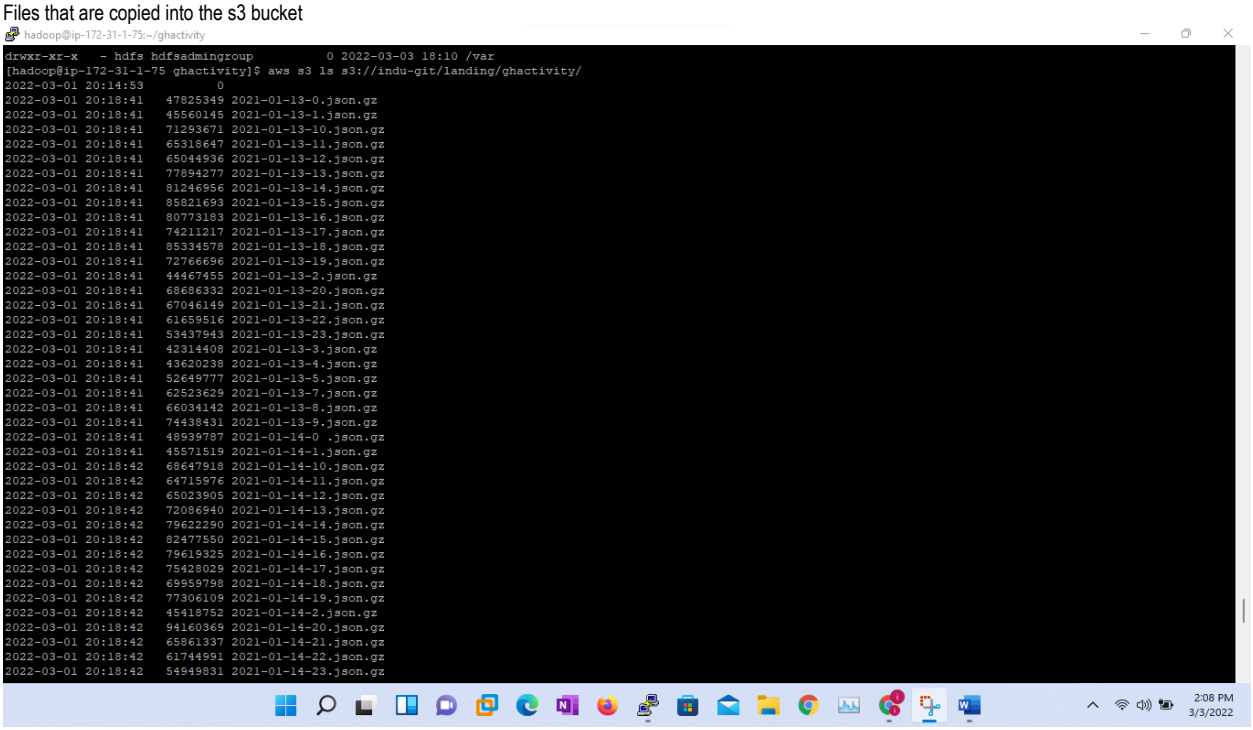


Copying the files from local to S3 Bucket using HDFS CLI on EMR Cluster

```
[hadoop@ip-172-31-1-75 ghactivity]$ hdfs dfs -copyFromLocal * s3://indu-git/landing/ghactivity
2022-03-03 20:02:14,549 INFO s3n.MultipartUploadOutputStream: close closed:false s3://indu-git/landing/ghactivity/2021-01-16-0.json.gz
2022-03-03 20:02:15,144 INFO s3n.MultipartUploadOutputStream: close closed:true s3://indu-git/landing/ghactivity/2021-01-16-0.json.gz
2022-03-03 20:02:15,340 INFO s3n.MultipartUploadOutputStream: close closed:false s3://indu-git/landing/ghactivity/2021-01-16-10.json.gz
2022-03-03 20:02:15,658 INFO s3n.MultipartUploadOutputStream: close closed:true s3://indu-git/landing/ghactivity/2021-01-16-10.json.gz
2022-03-03 20:02:15,854 INFO s3n.MultipartUploadOutputStream: close closed:false s3://indu-git/landing/ghactivity/2021-01-16-11.json.gz
2022-03-03 20:02:16,154 INFO s3n.MultipartUploadOutputStream: close closed:true s3://indu-git/landing/ghactivity/2021-01-16-11.json.gz
2022-03-03 20:02:16,342 INFO s3n.MultipartUploadOutputStream: close closed:false s3://indu-git/landing/ghactivity/2021-01-16-12.json.gz
2022-03-03 20:02:16,654 INFO s3n.MultipartUploadOutputStream: close closed:true s3://indu-git/landing/ghactivity/2021-01-16-12.json.gz
2022-03-03 20:02:16,823 INFO s3n.MultipartUploadOutputStream: close closed:false s3://indu-git/landing/ghactivity/2021-01-16-13.json.gz
2022-03-03 20:02:17,198 INFO s3n.MultipartUploadOutputStream: close closed:true s3://indu-git/landing/ghactivity/2021-01-16-13.json.gz
2022-03-03 20:02:17,380 INFO s3n.MultipartUploadOutputStream: close closed:false s3://indu-git/landing/ghactivity/2021-01-16-14.json.gz
2022-03-03 20:02:17,662 INFO s3n.MultipartUploadOutputStream: close closed:true s3://indu-git/landing/ghactivity/2021-01-16-14.json.gz
2022-03-03 20:02:17,847 INFO s3n.MultipartUploadOutputStream: close closed:false s3://indu-git/landing/ghactivity/2021-01-16-15.json.gz
2022-03-03 20:02:18,206 INFO s3n.MultipartUploadOutputStream: close closed:true s3://indu-git/landing/ghactivity/2021-01-16-15.json.gz
2022-03-03 20:02:18,405 INFO s3n.MultipartUploadOutputStream: close closed:false s3://indu-git/landing/ghactivity/2021-01-16-16.json.gz
2022-03-03 20:02:18,694 INFO s3n.MultipartUploadOutputStream: close closed:true s3://indu-git/landing/ghactivity/2021-01-16-16.json.gz
2022-03-03 20:02:18,874 INFO s3n.MultipartUploadOutputStream: close closed:false s3://indu-git/landing/ghactivity/2021-01-16-17.json.gz
2022-03-03 20:02:19,156 INFO s3n.MultipartUploadOutputStream: close closed:true s3://indu-git/landing/ghactivity/2021-01-16-17.json.gz
2022-03-03 20:02:19,348 INFO s3n.MultipartUploadOutputStream: close closed:false s3://indu-git/landing/ghactivity/2021-01-16-18.json.gz
2022-03-03 20:02:19,648 INFO s3n.MultipartUploadOutputStream: close closed:true s3://indu-git/landing/ghactivity/2021-01-16-18.json.gz
2022-03-03 20:02:19,826 INFO s3n.MultipartUploadOutputStream: close closed:false s3://indu-git/landing/ghactivity/2021-01-16-19.json.gz
2022-03-03 20:02:20,159 INFO s3n.MultipartUploadOutputStream: close closed:true s3://indu-git/landing/ghactivity/2021-01-16-19.json.gz
2022-03-03 20:02:20,318 INFO s3n.MultipartUploadOutputStream: close closed:false s3://indu-git/landing/ghactivity/2021-01-16-1.json.gz
2022-03-03 20:02:20,574 INFO s3n.MultipartUploadOutputStream: close closed:true s3://indu-git/landing/ghactivity/2021-01-16-1.json.gz
2022-03-03 20:02:20,736 INFO s3n.MultipartUploadOutputStream: close closed:false s3://indu-git/landing/ghactivity/2021-01-16-20.json.gz
2022-03-03 20:02:20,984 INFO s3n.MultipartUploadOutputStream: close closed:true s3://indu-git/landing/ghactivity/2021-01-16-20.json.gz
2022-03-03 20:02:21,141 INFO s3n.MultipartUploadOutputStream: close closed:false s3://indu-git/landing/ghactivity/2021-01-16-21.json.gz
2022-03-03 20:02:21,368 INFO s3n.MultipartUploadOutputStream: close closed:true s3://indu-git/landing/ghactivity/2021-01-16-21.json.gz
2022-03-03 20:02:21,517 INFO s3n.MultipartUploadOutputStream: close closed:false s3://indu-git/landing/ghactivity/2021-01-16-22.json.gz
2022-03-03 20:02:21,760 INFO s3n.MultipartUploadOutputStream: close closed:true s3://indu-git/landing/ghactivity/2021-01-16-22.json.gz
2022-03-03 20:02:21,898 INFO s3n.MultipartUploadOutputStream: close closed:false s3://indu-git/landing/ghactivity/2021-01-16-23.json.gz
2022-03-03 20:02:22,182 INFO s3n.MultipartUploadOutputStream: close closed:true s3://indu-git/landing/ghactivity/2021-01-16-23.json.gz
2022-03-03 20:02:22,333 INFO s3n.MultipartUploadOutputStream: close closed:false s3://indu-git/landing/ghactivity/2021-01-16-2.json.gz
2022-03-03 20:02:22,597 INFO s3n.MultipartUploadOutputStream: close closed:true s3://indu-git/landing/ghactivity/2021-01-16-2.json.gz
2022-03-03 20:02:22,736 INFO s3n.MultipartUploadOutputStream: close closed:false s3://indu-git/landing/ghactivity/2021-01-16-3.json.gz
2022-03-03 20:02:22,984 INFO s3n.MultipartUploadOutputStream: close closed:true s3://indu-git/landing/ghactivity/2021-01-16-3.json.gz
```



Files that are copied into the s3 bucket



Code copied from local to EMR cluster using SCP.

scp -i EMRClusterkeypair.pem itv-ghactivity.zip hadoop@ec2-18-223-172-40.us-east-2.compute.amazonaws.com:~

scp -i EMRClusterkeypair.pem app.py hadoop@ec2-18-223-172-40.us-east-2.compute.amazonaws.com:~

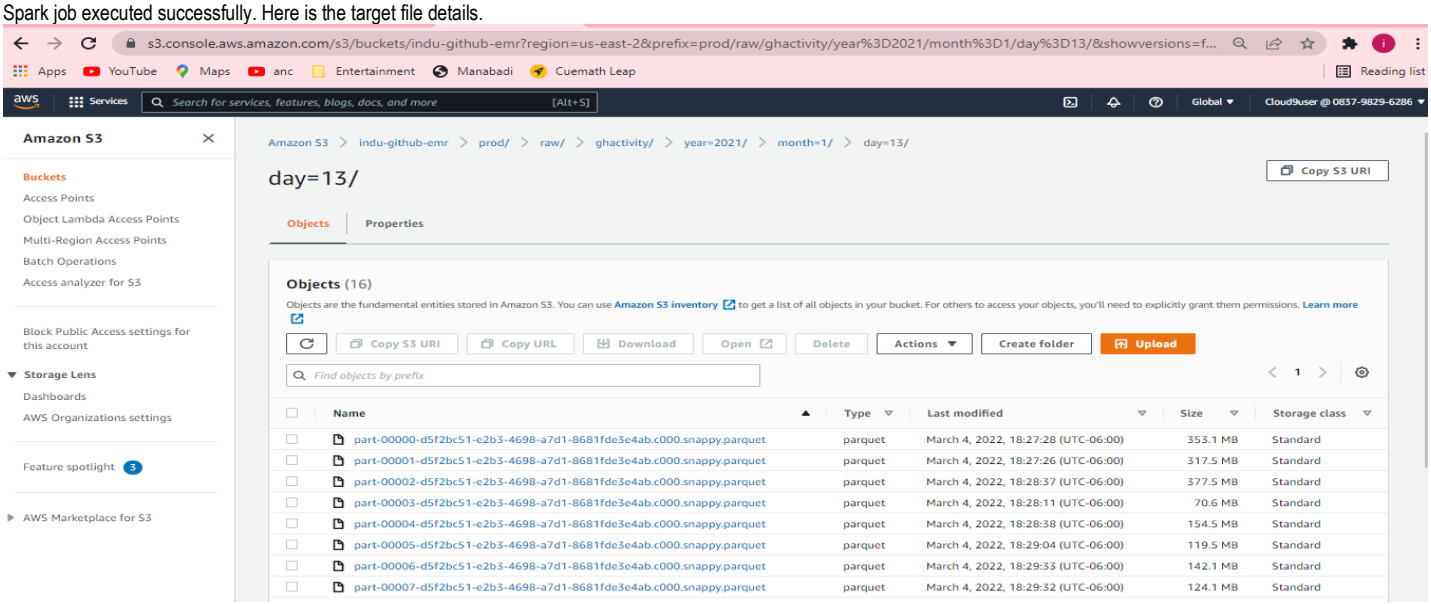
I. Running spark job to read the files from above mentioned location and write into parquet format. Also setup environment variables like below:

```
[hadoop@ip-172-31-2-100 itv-ghactivity]$ history | grep export
    3  export export
    4  export ENVIRON=PROD
    5  export SRD_DIR=S3://indu-github-emr/prod/landing/ghactivity
    6  export SRC_FILE_FORMAT=json
    7  export TRT_DIR=s3://indu-github-emmmmr/prod/raw/ghactivity/
    8  export SRD_DIR=S3://indu-github-emr/prod/landing/ghactivity/
    9  export TGT_FILE_FORMAT=parquet
   10  export SRC_FILE_PATTERN=2021-01-13
   11  history|grep export
   22  history|grep export
   23  export TGT_DIR=s3://indu-github-emmmmr/prod/raw/ghactivity/
   27  export TGT_DIR=s3://indu-github-emr/prod/raw/ghactivity/
   42  history | grep export
   47  history | grep export
[hadoop@ip-172-31-2-100 itv-ghactivity]$
```

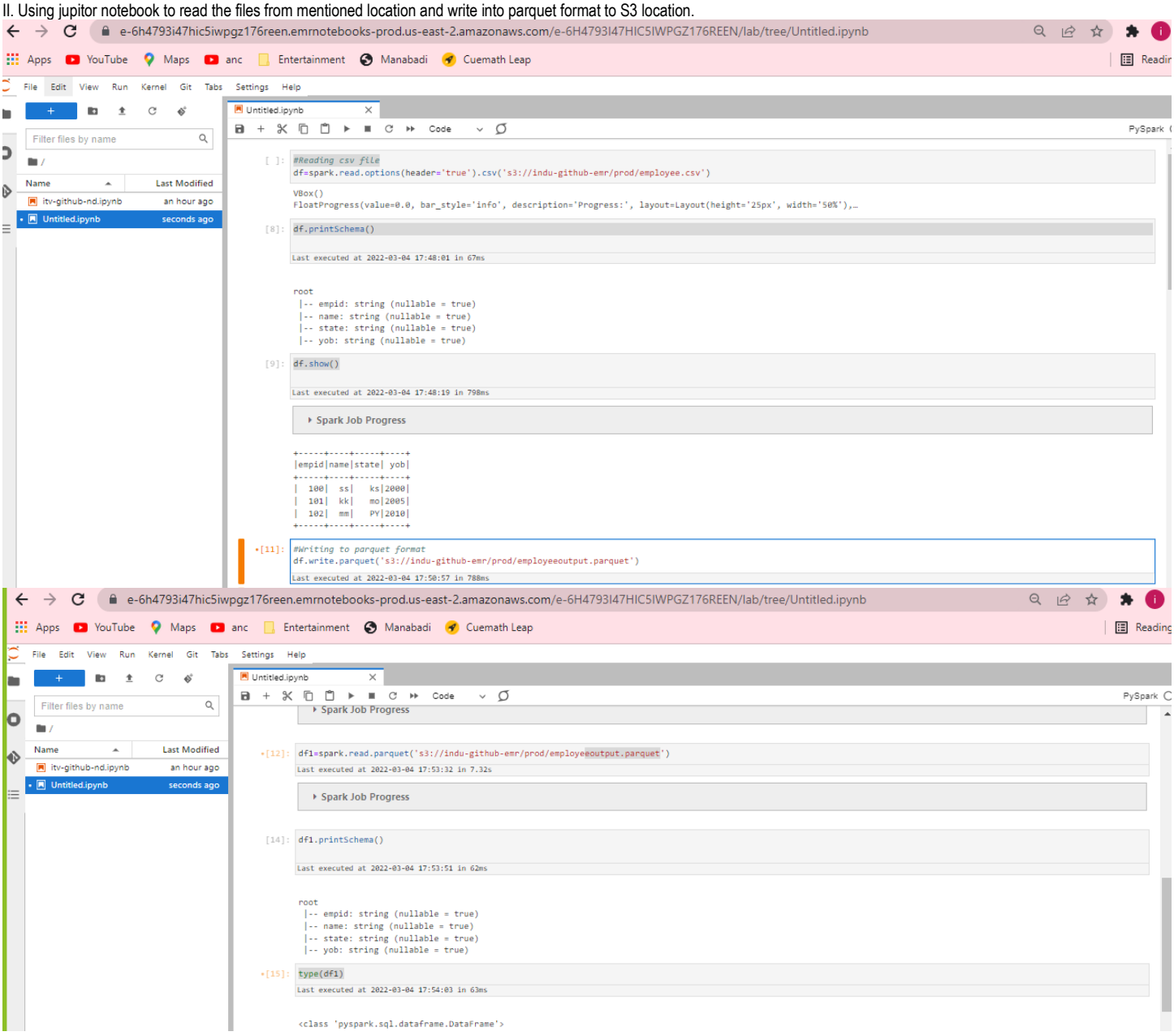
[hadoop@ip-172-31-2-100 itv-ghactivity]$ spark-submit --master yarn --py-files itv-ghactivity.zip app.py

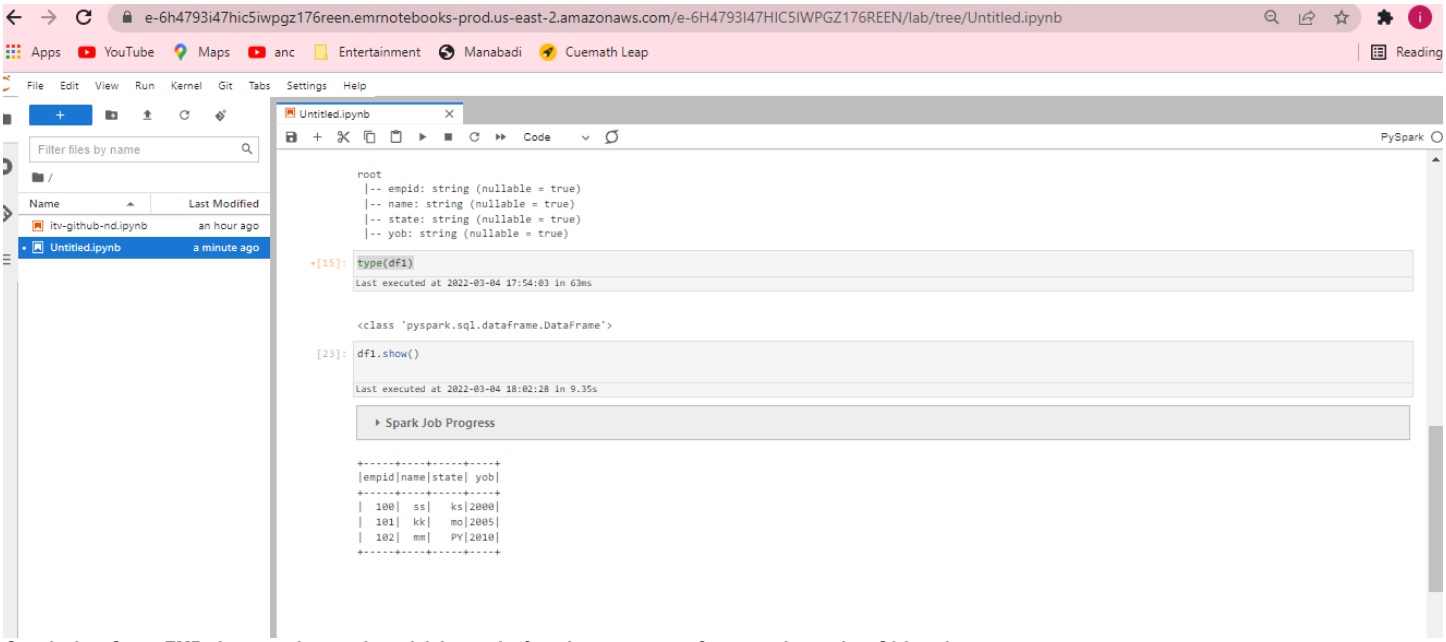
```
hadoop@ip-172-31-2-100:~/itv-ghactivity
bytes) taskResourceAssignments Map()
22/03/05 00:30:00 INFO TaskSetManager: Finished task 9.0 in stage 1.0 (TID 28) in 25009 ms on ip-172-31-5-189.us-east-2.compute.internal (executor 1) (10/16)
22/03/05 00:30:24 INFO TaskSetManager: Starting task 12.0 in stage 1.0 (TID 31) (ip-172-31-5-189.us-east-2.compute.internal, executor 1, partition 12, RACK_LOCAL, 5397
bytes) taskResourceAssignments Map()
22/03/05 00:30:24 INFO TaskSetManager: Finished task 10.0 in stage 1.0 (TID 29) in 28798 ms on ip-172-31-5-189.us-east-2.compute.internal (executor 1) (11/16)
22/03/05 00:30:32 INFO TaskSetManager: Starting task 13.0 in stage 1.0 (TID 32) (ip-172-31-5-189.us-east-2.compute.internal, executor 1, partition 13, RACK_LOCAL, 5397
bytes) taskResourceAssignments Map()
22/03/05 00:30:32 INFO TaskSetManager: Finished task 11.0 in stage 1.0 (TID 30) in 32037 ms on ip-172-31-5-189.us-east-2.compute.internal (executor 1) (12/16)
22/03/05 00:30:50 INFO TaskSetManager: Starting task 14.0 in stage 1.0 (TID 33) (ip-172-31-5-189.us-east-2.compute.internal, executor 1, partition 14, RACK_LOCAL, 5397
bytes) taskResourceAssignments Map()
22/03/05 00:30:50 INFO TaskSetManager: Finished task 12.0 in stage 1.0 (TID 31) in 25185 ms on ip-172-31-5-189.us-east-2.compute.internal (executor 1) (13/16)
22/03/05 00:30:56 INFO TaskSetManager: Starting task 15.0 in stage 1.0 (TID 34) (ip-172-31-5-189.us-east-2.compute.internal, executor 1, partition 15, RACK_LOCAL, 5396
bytes) taskResourceAssignments Map()
22/03/05 00:30:56 INFO TaskSetManager: Finished task 13.0 in stage 1.0 (TID 32) in 24048 ms on ip-172-31-5-189.us-east-2.compute.internal (executor 1) (14/16)
22/03/05 00:31:13 INFO TaskSetManager: Finished task 14.0 in stage 1.0 (TID 33) in 23042 ms on ip-172-31-5-189.us-east-2.compute.internal (executor 1) (15/16)
22/03/05 00:31:18 INFO TaskSetManager: Finished task 15.0 in stage 1.0 (TID 34) in 21706 ms on ip-172-31-5-189.us-east-2.compute.internal (executor 1) (16/16)
22/03/05 00:31:18 INFO YarnScheduler: Removed TaskSet 1.0, whose tasks have all completed, from pool
22/03/05 00:31:18 INFO DAGScheduler: ResultStage 1 (save at NativeMethodAccessorImpl.java:0) finished in 279.637 s
22/03/05 00:31:18 INFO DAGScheduler: Job 1 is finished. Cancelling potential speculative or zombie tasks for this job
22/03/05 00:31:18 INFO YarnScheduler: Killing all running tasks in stage 1: Stage finished
22/03/05 00:31:18 INFO DAGScheduler: Job 1 finished: save at NativeMethodAccessorImpl.java:0, took 279.648022 s
22/03/05 00:31:18 INFO MultipartUploadOutputStream: close closed:false s3://indu-github-emr/prod/raw/ghactivity/_SUCCESS
22/03/05 00:31:18 INFO FileFormatWriter: Write Job cd4afa79-bbeb-4764-bd83-48fd891d8556 committed.
22/03/05 00:31:18 INFO FileFormatWriter: Finished processing stats for write job cd4afa79-bbeb-4764-bd83-48fd891d8556.
22/03/05 00:31:18 INFO SparkContext: Invoking stop() from shutdown hook
22/03/05 00:31:18 INFO AbstractConnector: Stopped Spark@668b73ca{HTTP/1.1, (http/1.1)}{0.0.0.0:4041}
22/03/05 00:31:18 INFO SparkUI: Stopped Spark web UI at http://ip-172-31-2-100.us-east-2.compute.internal:4041
22/03/05 00:31:18 INFO YarnClientSchedulerBackend: Interrupting monitor thread
22/03/05 00:31:18 INFO YarnClientSchedulerBackend: Shutting down all executors
22/03/05 00:31:18 INFO YarnSchedulerBackend$YarnDriverEndpoint: Asking each executor to shut down
22/03/05 00:31:18 INFO YarnClientSchedulerBackend: YARN client scheduler backend Stopped
22/03/05 00:31:18 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
22/03/05 00:31:18 INFO MemoryStore: MemoryStore cleared
22/03/05 00:31:18 INFO BlockManager: BlockManager stopped
22/03/05 00:31:18 INFO BlockManagerMaster: BlockManagerMaster stopped
22/03/05 00:31:18 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
22/03/05 00:31:18 INFO SparkContext: Successfully stopped SparkContext
22/03/05 00:31:18 INFO ShutdownHookManager: Shutdown hook called
22/03/05 00:31:18 INFO ShutdownHookManager: Deleting directory /mnt/tmp/spark-95a4f3c1-92b1-4a51-957c-ff2fc1185d5f
22/03/05 00:31:18 INFO ShutdownHookManager: Deleting directory /mnt/tmp/spark-95a4f3c1-92b1-4a51-957c-ff2fc1185d5f/pyspark-1508ee8e-76ee-4d24-95a6-937075a91272
22/03/05 00:31:18 INFO ShutdownHookManager: Deleting directory /mnt/tmp/spark-7bf3dd8a-1e21-4863-9d7f-a5447c834d63
```

Spark job executed successfully. Here is the target file details.



II. Using jupitor notebook to read the files from mentioned location and write into parquet format to S3 location.





Conclusion: Setup EMR cluster and created spark job to write from json to parquet format and store into S3 location.

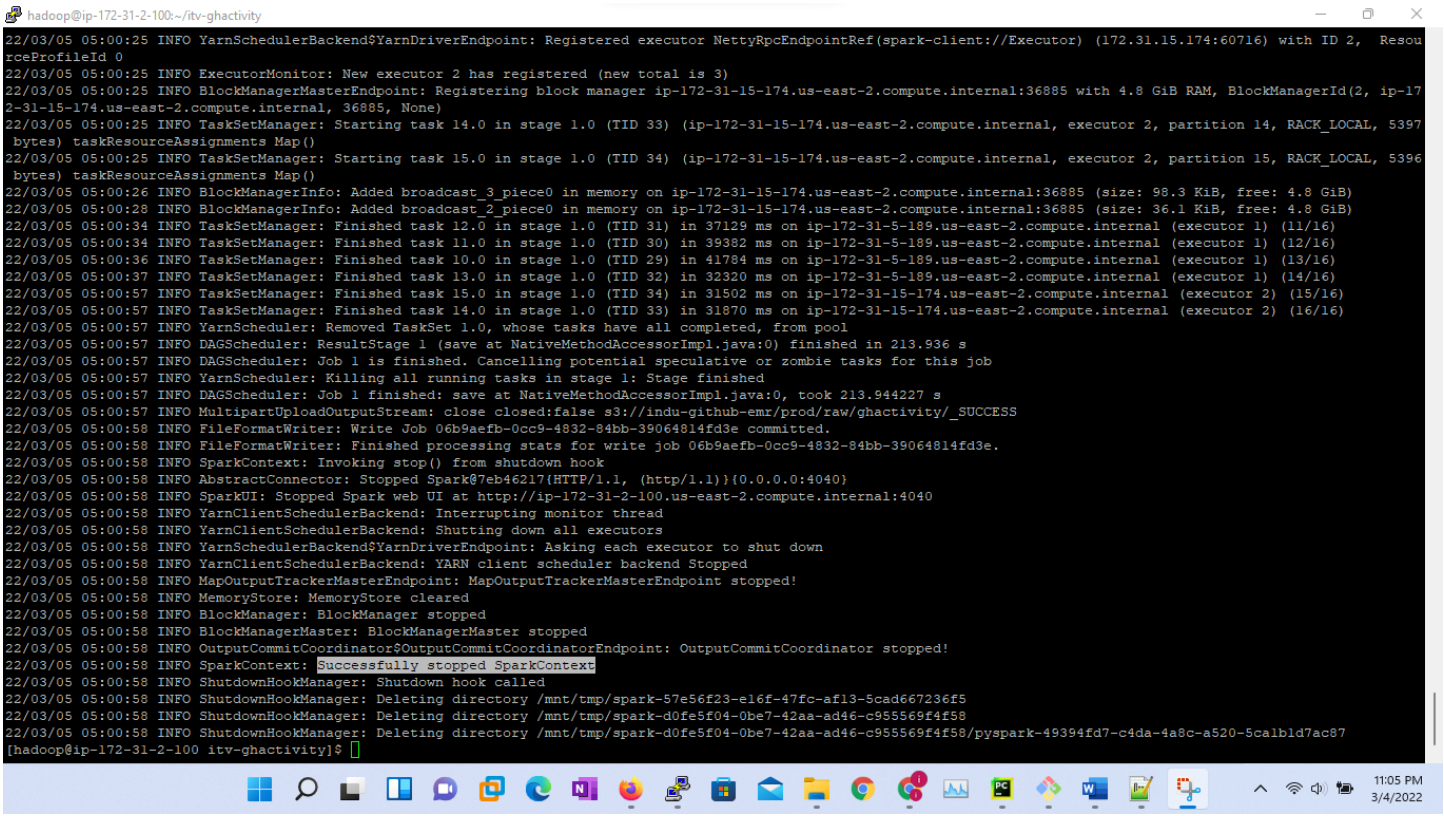
Deleted the data in s3 bucket :

aws s3 rm s3://indu-github-emr/prod/raw/ghactivity/ --recursive

Running Spark Application Using Cluster mode on AWS EMR Cluster

In the cluster mode we need the pass the environmental variables in the spark-submit command itself then it runs successfully.

```
spark-submit \
        --master yarn \
        --conf "spark.yarn.appMasterEnv.ENVIRON=PROD" \
        --conf "spark.yarn.appMasterEnv.SRD_DIR=S3://indu-github-emr/prod/landing/ghactivity/" \
        --conf "spark.yarn.appMasterEnv.SRC_FILE_FORMAT=json" \
        --conf "spark.yarn.appMasterEnv.TGT_DIR=s3://indu-github-emr/prod/raw/ghactivity/" \
        --conf "spark.yarn.appMasterEnv.TGT_FILE_FORMAT=parquet" \
        --conf "spark.yarn.appMasterEnv.SRC_FILE_PATTERN=2021-01-13" \
        --py-files itv-ghactivity.zip \
        app.py
```

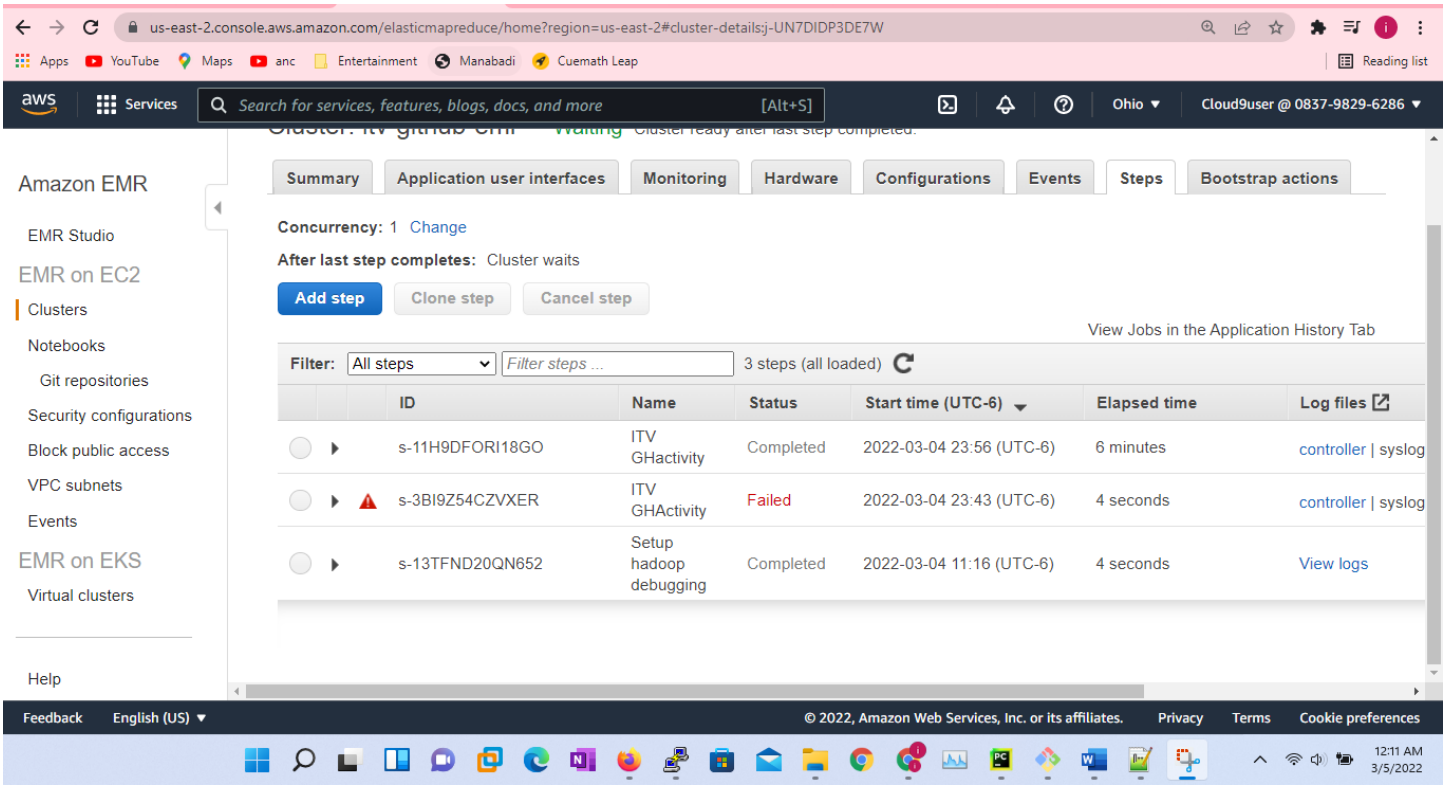


Running Spark Application as AWS EMR steps in Cluster mode

In the step we need to pass the arguments

--conf spark.yarn.appMasterEnv.ENVIRON=PROD --conf spark.yarn.appMasterEnv.SRD_DIR=S3://indu-github-emr/prod/landing/ghactivity/ --conf spark.yarn.appMasterEnv.SRC_FILE_FORMAT=json --conf spark.yarn.appMasterEnv.TGT_DIR=s3://indu-github-emr/prod/raw/ghactivity/ --conf spark.yarn.appMasterEnv.TGT_FILE_FORMAT=parquet --conf spark.yarn.appMasterEnv.SRC_FILE_PATTERN=2021-01-14 --py-files s3://indu-github-emr/app/itv-ghactivity.zip

And specify jar location is S3;// indu-github-emr/app/app.py


us-east-2.console.aws.amazon.com/elasticmapreduce/home?region=us-east-2#cluster-details:j-UN7DIDP3DE7W
Apps
YouTube
Maps
anc
Entertainment
Manabadi
Cuemath Leap
Reading list
aws
Services
Search for services, features, blogs, docs, and more
[Alt+S]
Ohio
Cloud9user @ 0837-9829-6286
Amazon EMR
EMR Studio
EMR on EC2
Clusters
Notebooks
Git repositories
Security configurations
Block public access
VPC subnets
Events
EMR on EKS
Virtual clusters
Help
Summary
Application user interfaces
Monitoring
Hardware
Configurations
Events
Steps
Bootstrap actions
Concurrency: 1 Change
After last step completes: Cluster waits
Add step
Clone step
Cancel step
View Jobs in the Application History Tab
Filter: All steps
Filter steps ...
3 steps (all loaded)
ID
Name
Status
Start time (UTC-6)
Elapsed time
Log files
s-11H9DFORI18GO
ITV GHactivity
Completed
2022-03-04 23:56 (UTC-6)
6 minutes
controller | syslog
s-3BI9Z54CZVXER
ITV GHActivity
Failed
2022-03-04 23:43 (UTC-6)
4 seconds
controller | syslog
s-13TFND20QN652
Setup hadoop debugging
Completed
2022-03-04 11:16 (UTC-6)
4 seconds
View logs
Feedback
English (US)
© 2022, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences
12:11 AM
3/5/2022