

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

-
- The screenshot shows the AWS Management Console interface. The left sidebar contains navigation links for various AWS services: Elastic Block Store, Network & Security, Load Balancing, and Auto Scaling. The main content area is titled 'Key pairs (1/2)' and shows a table of key pairs. The table has columns for Name, Type, and Fingerprint. Two key pairs are listed: 'EMRClusterkeypair' (Type: rsa) and 'awskey' (Type: rsa). The 'EMRClusterkeypair' is selected, and its details are shown below the table. The details include the key pair name, type, and fingerprint. The bottom of the screenshot shows the Windows taskbar with various application icons and the system clock.
- | Name | Type | Fingerprint |
|-------------------|------|--|
| EMRClusterkeypair | rsa | 2d:80:72:47:76:3e:03:1d:8b:62:e7:b3:7... |
| awskey | rsa | b8:4a:54:e9:91:25:0f:68:30:64:37:61:8... |

Using EMR cluster ipaddress started the terminal using SSH

Figure 2 Connected EMR master node using SSH

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

```
hadoop@ip-172-31-1-75:~  
E::::E      EEEEE M::::M      MMM      M::::M      R::::R      R::::R  
EE:::::EEEEEEEEE::::E M::::M      M::::M      R::::R      R::::R  
E:::::EEEEEEEEE::::E M::::M      M::::M      R::::R      R::::R  
EEEEEEEEE EEEEEEEEEE MMMMMMM      MMMMMMM      RRRRRRR      RRRRRR  
  
[hadoop@ip-172-31-1-75 ~]$ aws s3 ls  
2022-02-26 23:23:41 aws-glue-scripts-083798296286-us-east-2  
2022-02-26 23:23:43 aws-glue-temporary-083798296286-us-east-2  
2022-03-03 04:49:59 aws-logs-083798296286-us-east-2  
2022-02-22 19:45:21 indu-git  
2022-02-20 02:44:45 indu-retail  
2022-02-20 02:43:39 indu-retail-copy  
2022-02-26 22:15:01 itv-flightst16  
2022-02-26 21:48:46 itvathena1  
[hadoop@ip-172-31-1-75 ~]$ aws s3 ls s3://indu-git  
PRE ghactivity/  
PRE landing/  
PRE raw/  
PRE sandbox/  
PRE spark-logs/  
2022-02-23 04:23:37 60178683 2022-02-22-0.json.gz  
2022-03-02 05:27:56 0 ghactivity_folder$  
[hadoop@ip-172-31-1-75 ~]$ aws s3 ls s3://indu-git/landing  
PRE landing/  
[hadoop@ip-172-31-1-75 ~]$ aws s3 ls s3://indu-git/landing/  
PRE ghactivity/  
2022-02-22 19:46:00 0  
[hadoop@ip-172-31-1-75 ~]$ aws s3 ls s3://indu-git/landing/PRE ghactivity  
  
Unknown options: ghactivity  
[hadoop@ip-172-31-1-75 ~]$ aws s3 ls s3://indu-git/landing/ghactivity  
PRE ghactivity/  
[hadoop@ip-172-31-1-75 ~]$ aws s3 ls s3://indu-git/landing/ghactivity/  
2022-03-01 20:14:53 0  
2022-03-01 20:18:41 47825349 2021-01-13-0.json.gz  
2022-03-01 20:18:41 45560145 2021-01-13-1.json.gz  
2022-03-01 20:18:41 71293671 2021-01-13-10.json.gz  
2022-03-01 20:18:41 65318647 2021-01-13-11.json.gz  
2022-03-01 20:18:41 65044936 2021-01-13-12.json.gz  
2022-03-01 20:18:41 77894277 2021-01-13-13.json.gz  
2022-03-01 20:18:41 81246956 2021-01-13-14.json.gz  
2022-03-01 20:18:41 85821693 2021-01-13-15.json.gz  
2022-03-01 20:18:41 80773183 2021-01-13-16.json.gz
```

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

```
hadoop@ip-172-31-1-75:~  
[hadoop@ip-172-31-1-75 ~]$ hdfs dfs -ls  
[hadoop@ip-172-31-1-75 ~]$ hdfs dfs -ls /  
Found 3 items  
drwxrwxrwt - hdfs hdfsadmingroup 0 2022-03-03 18:10 /tmp  
drwxr-xr-x - hdfs hdfsadmingroup 0 2022-03-03 18:10 /user  
drwxr-xr-x - hdfs hdfsadmingroup 0 2022-03-03 18:10 /var  
[hadoop@ip-172-31-1-75 ~]$ hadoop fs -ls /  
Found 3 items  
drwxrwxrwt - hdfs hdfsadmingroup 0 2022-03-03 18:10 /tmp  
drwxr-xr-x - hdfs hdfsadmingroup 0 2022-03-03 18:10 /user  
drwxr-xr-x - hdfs hdfsadmingroup 0 2022-03-03 18:10 /var  
[hadoop@ip-172-31-1-75 ~]$ hdfs dfs -ls s3://indu-git/  
Found 6 items  
-rw-rw-rw- 1 hadoop hadoop 60178683 2022-02-23 04:23 s3://indu-git/2022-02-22-0.json.gz  
drwxrwxrwx - hadoop hadoop 0 1970-01-01 00:00 s3://indu-git/ghactivity  
drwxrwxrwx - hadoop hadoop 0 1970-01-01 00:00 s3://indu-git/landing  
drwxrwxrwx - hadoop hadoop 0 1970-01-01 00:00 s3://indu-git/raw  
drwxrwxrwx - hadoop hadoop 0 1970-01-01 00:00 s3://indu-git/sandbox  
drwxrwxrwx - hadoop hadoop 0 1970-01-01 00:00 s3://indu-git/spark-logs  
[hadoop@ip-172-31-1-75 ~]$ hdfs dfs -ls s3://indu-git/landing/  
Found 1 items  
drwxrwxrwx - hadoop hadoop 0 1970-01-01 00:00 s3://indu-git/landing/ghactivity  
[hadoop@ip-172-31-1-75 ~]$ hdfs dfs -ls s3://indu-git/landing/ghactivity  
Found 71 items  
-rw-rw-rw- 1 hadoop hadoop 47825349 2022-03-01 20:18 s3://indu-git/landing/ghactivity/2021-01-13-0.json.gz  
-rw-rw-rw- 1 hadoop hadoop 45560145 2022-03-01 20:18 s3://indu-git/landing/ghactivity/2021-01-13-1.json.gz  
-rw-rw-rw- 1 hadoop hadoop 71293671 2022-03-01 20:18 s3://indu-git/landing/ghactivity/2021-01-13-10.json.gz  
-rw-rw-rw- 1 hadoop hadoop 65318647 2022-03-01 20:18 s3://indu-git/landing/ghactivity/2021-01-13-11.json.gz  
-rw-rw-rw- 1 hadoop hadoop 65044936 2022-03-01 20:18 s3://indu-git/landing/ghactivity/2021-01-13-12.json.gz  
-rw-rw-rw- 1 hadoop hadoop 77894277 2022-03-01 20:18 s3://indu-git/landing/ghactivity/2021-01-13-13.json.gz  
-rw-rw-rw- 1 hadoop hadoop 81246956 2022-03-01 20:18 s3://indu-git/landing/ghactivity/2021-01-13-14.json.gz  
-rw-rw-rw- 1 hadoop hadoop 85821693 2022-03-01 20:18 s3://indu-git/landing/ghactivity/2021-01-13-15.json.gz  
-rw-rw-rw- 1 hadoop hadoop 80773183 2022-03-01 20:18 s3://indu-git/landing/ghactivity/2021-01-13-16.json.gz  
-rw-rw-rw- 1 hadoop hadoop 74211217 2022-03-01 20:18 s3://indu-git/landing/ghactivity/2021-01-13-17.json.gz  
-rw-rw-rw- 1 hadoop hadoop 85334578 2022-03-01 20:18 s3://indu-git/landing/ghactivity/2021-01-13-18.json.gz  
-rw-rw-rw- 1 hadoop hadoop 72766696 2022-03-01 20:18 s3://indu-git/landing/ghactivity/2021-01-13-19.json.gz  
-rw-rw-rw- 1 hadoop hadoop 44467455 2022-03-01 20:18 s3://indu-git/landing/ghactivity/2021-01-13-2.json.gz  
-rw-rw-rw- 1 hadoop hadoop 68686332 2022-03-01 20:18 s3://indu-git/landing/ghactivity/2021-01-13-20.json.gz  
-rw-rw-rw- 1 hadoop hadoop 67046149 2022-03-01 20:18 s3://indu-git/landing/ghactivity/2021-01-13-21.json.gz  
-rw-rw-rw- 1 hadoop hadoop 61659516 2022-03-01 20:18 s3://indu-git/landing/ghactivity/2021-01-13-22.json.gz  
-rw-rw-rw- 1 hadoop hadoop 53437943 2022-03-01 20:18 s3://indu-git/landing/ghactivity/2021-01-13-23.json.gz  
-rw-rw-rw- 1 hadoop hadoop 43314408 2022-03-01 20:18 s3://indu-git/landing/ghactivity/2021-01-13-3.json.gz  
-rw-rw-rw- 1 hadoop hadoop 43620238 2022-03-01 20:18 s3://indu-git/landing/ghactivity/2021-01-13-4.json.gz
```

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 20:02:17,196 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

```
hadoop@ip-172-31-1-75:~/ghactivity
drwxr-xr-x - hdfs hdfsadmingroup 0 2022-03-03 18:10 /var
[hadoop@ip-172-31-1-75 ghactivity]$ aws ls s3://indu-git/landing/ghactivity/
2022-03-01 20:14:53 0
2022-03-01 20:18:41 47825349 2021-01-13-0.json.gz
2022-03-01 20:18:41 45560145 2021-01-13-1.json.gz
2022-03-01 20:18:41 71293671 2021-01-13-10.json.gz
2022-03-01 20:18:41 65318647 2021-01-13-11.json.gz
2022-03-01 20:18:41 65044936 2021-01-13-12.json.gz
2022-03-01 20:18:41 77894277 2021-01-13-13.json.gz
2022-03-01 20:18:41 81246956 2021-01-13-14.json.gz
2022-03-01 20:18:41 85821693 2021-01-13-15.json.gz
2022-03-01 20:18:41 80773183 2021-01-13-16.json.gz
2022-03-01 20:18:41 74211217 2021-01-13-17.json.gz
2022-03-01 20:18:41 85334578 2021-01-13-18.json.gz
2022-03-01 20:18:41 72766866 2021-01-13-19.json.gz
2022-03-01 20:18:41 44467455 2021-01-13-2.json.gz
2022-03-01 20:18:41 65686332 2021-01-13-20.json.gz
2022-03-01 20:18:41 67046149 2021-01-13-21.json.gz
2022-03-01 20:18:41 61659516 2021-01-13-22.json.gz
2022-03-01 20:18:41 53437943 2021-01-13-23.json.gz
2022-03-01 20:18:41 42314408 2021-01-13-3.json.gz
2022-03-01 20:18:41 43620238 2021-01-13-4.json.gz
2022-03-01 20:18:41 52649777 2021-01-13-5.json.gz
2022-03-01 20:18:41 62523629 2021-01-13-7.json.gz
2022-03-01 20:18:41 66034142 2021-01-13-8.json.gz
2022-03-01 20:18:41 74438431 2021-01-13-9.json.gz
2022-03-01 20:18:41 8939787 2021-01-14-0.json.gz
2022-03-01 20:18:41 45571519 2021-01-14-1.json.gz
2022-03-01 20:18:42 68647918 2021-01-14-10.json.gz
2022-03-01 20:18:42 64715976 2021-01-14-11.json.gz
2022-03-01 20:18:42 65023905 2021-01-14-12.json.gz
2022-03-01 20:18:42 72086940 2021-01-14-13.json.gz
2022-03-01 20:18:42 79622290 2021-01-14-14.json.gz
2022-03-01 20:18:42 82477550 2021-01-14-15.json.gz
2022-03-01 20:18:42 79619325 2021-01-14-16.json.gz
2022-03-01 20:18:42 75428029 2021-01-14-17.json.gz
2022-03-01 20:18:42 69959798 2021-01-14-18.json.gz
2022-03-01 20:18:42 77306109 2021-01-14-19.json.gz
2022-03-01 20:18:42 45418752 2021-01-14-2.json.gz
2022-03-01 20:18:42 94160369 2021-01-14-20.json.gz
2022-03-01 20:18:42 65861337 2021-01-14-21.json.gz
2022-03-01 20:18:42 61744991 2021-01-14-22.json.gz
2022-03-01 20:18:42 54949831 2021-01-14-23.json.gz
```

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:~
```

```
hadoop@ip-172-31-2-100:~/itv-ghactivity
[hadoop@ip-172-31-2-100 itv-ghactivity]$ ll
total 28
-rw-rw-r-- 1 hadoop hadoop 1573 Mar  5 00:25 app.py
-rw-r--r-- 1 hadoop hadoop 2589 Mar  4 18:05 itv-ghactivity.zip
-rw-rw-r-- 1 hadoop hadoop 120 Feb 24 00:24 new.py
-rw-rw-r-- 1 hadoop hadoop 255 Feb 24 06:26 process.py
drwxrwxr-x 2 hadoop hadoop 118 Mar  4 23:12 pycache
-rw-rw-r-- 1 hadoop hadoop 178 Mar  4 06:56 read.py
-rw-rw-r-- 1 hadoop hadoop 343 Mar  4 17:09 util.py
-rw-rw-r-- 1 hadoop hadoop 326 Mar  4 16:49 write.py
[hadoop@ip-172-31-2-100 itv-ghactivity]$
```

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-emr/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-emr/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
```

```
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://1.1.1.1:4041))
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-ff2fcl185d5f
22/03/05 00:31:18 INFO ShutdownHookManager: Deleting directory /mnt/tmp/spark-95a4f3c1-92b1-4a51-957c-ff2fcl185d5f/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.

Amazon S3 console screenshot showing the details of the Spark job output files. The console displays the path: `s3.console.aws.amazon.com/s3/buckets/indu-github-emr?region=us-east-2&prefix=prod/raw/ghactivity/year%3D2021/month%3D1/day%3D13/&showversions=f...`. The left sidebar shows the navigation menu with options like Buckets, Access Points, and Storage Lens. The main content area shows the path: `Amazon S3 > indu-github-emr > prod/ > raw/ > ghactivity/ > year=2021/ > month=1/ > day=13/`. The file list shows 16 objects, all of type `parquet`, with names like `part-00000-d5f2bc51-e2b3-4698-a7d1-8681fde3e4ab.c000.snappy.parquet`. The table includes columns for Name, Type, Last modified, Size, and Storage class.

Name	Type	Last modified	Size	Storage class
part-00000-d5f2bc51-e2b3-4698-a7d1-8681fde3e4ab.c000.snappy.parquet	parquet	March 4, 2022, 18:27:28 (UTC-06:00)	353.1 MB	Standard
part-00001-d5f2bc51-e2b3-4698-a7d1-8681fde3e4ab.c000.snappy.parquet	parquet	March 4, 2022, 18:27:26 (UTC-06:00)	317.5 MB	Standard
part-00002-d5f2bc51-e2b3-4698-a7d1-8681fde3e4ab.c000.snappy.parquet	parquet	March 4, 2022, 18:28:37 (UTC-06:00)	377.5 MB	Standard
part-00003-d5f2bc51-e2b3-4698-a7d1-8681fde3e4ab.c000.snappy.parquet	parquet	March 4, 2022, 18:28:11 (UTC-06:00)	70.6 MB	Standard
part-00004-d5f2bc51-e2b3-4698-a7d1-8681fde3e4ab.c000.snappy.parquet	parquet	March 4, 2022, 18:28:38 (UTC-06:00)	154.5 MB	Standard
part-00005-d5f2bc51-e2b3-4698-a7d1-8681fde3e4ab.c000.snappy.parquet	parquet	March 4, 2022, 18:29:04 (UTC-06:00)	119.5 MB	Standard
part-00006-d5f2bc51-e2b3-4698-a7d1-8681fde3e4ab.c000.snappy.parquet	parquet	March 4, 2022, 18:29:33 (UTC-06:00)	142.1 MB	Standard
part-00007-d5f2bc51-e2b3-4698-a7d1-8681fde3e4ab.c000.snappy.parquet	parquet	March 4, 2022, 18:29:32 (UTC-06:00)	124.1 MB	Standard

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

The image displays two sequential screenshots of a Jupyter Notebook interface, likely running on an Amazon EMR instance. The browser address bar shows the URL: `e-6h4793i47hic5iwpgz176green.emrnotebooks-prod.us-east-2.amazonaws.com/e-6H4793I47HIC5IWPgz176GREEN/lab/tree/Untitled.ipynb`.

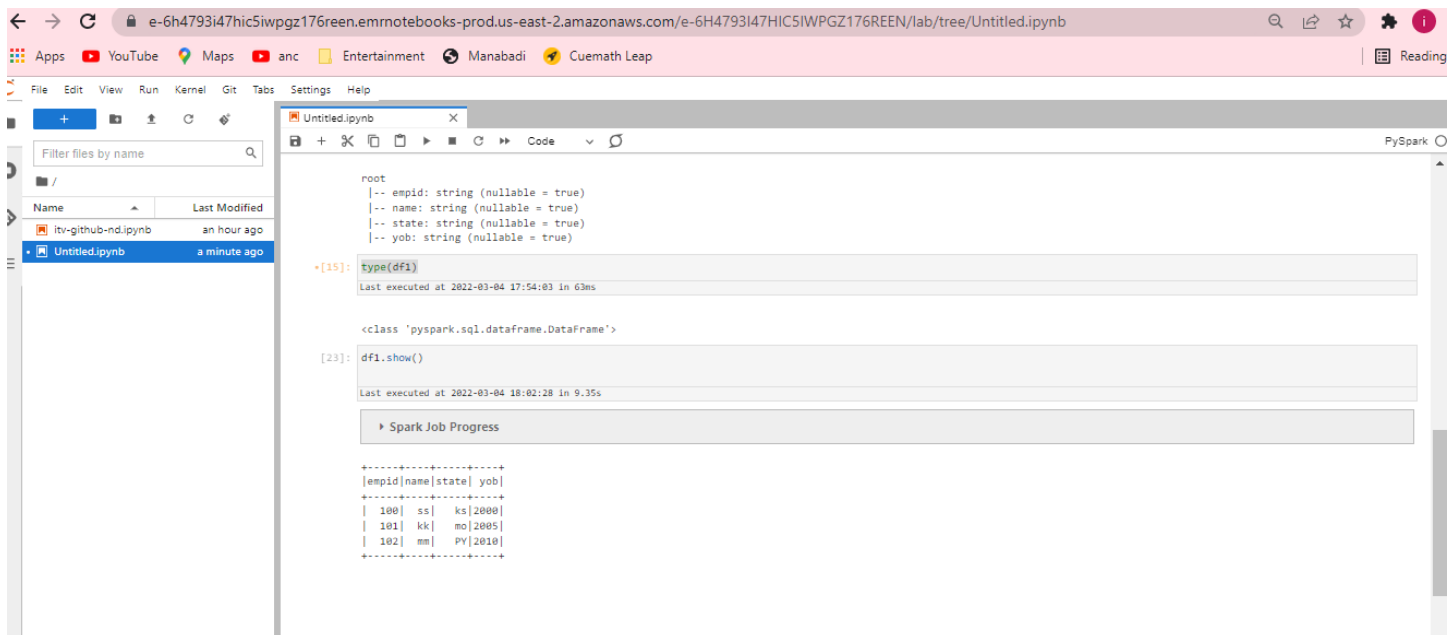
Top Screenshot: The notebook is titled "Untitled.ipynb". The code cell [8] shows the following steps:

- Reading a CSV file from S3: `df=spark.read.options(header='true').csv('s3://indu-github-emr/prod/employee.csv')`
- Displaying the schema: `df.printSchema()`
- The output shows the schema: `root |-- empid: string (nullable = true) |-- name: string (nullable = true) |-- state: string (nullable = true) |-- yob: string (nullable = true)`
- Displaying the data: `df.show()`
- The output shows a table with 3 rows and 4 columns:

empid	name	state	yob
100	ss	Ks	2000
101	kk	mo	2005
102	em	PV	2010
- Writing the data to Parquet format: `df.write.parquet('s3://indu-github-emr/prod/employeeoutput.parquet')`

Bottom Screenshot: The notebook continues with the following steps:

- Reading the Parquet file: `df1=spark.read.parquet('s3://indu-github-emr/prod/employeeoutput.parquet')`
- Displaying the schema: `df1.printSchema()`
- The output shows the schema: `root |-- empid: string (nullable = true) |-- name: string (nullable = true) |-- state: string (nullable = true) |-- yob: string (nullable = true)`
- Displaying the type of the dataframe: `type(df1)`
- The output shows the type: `<class 'pyspark.sql.dataframe.DataFrame'>`



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 to 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

```



```
hadoop@ip-172-31-2-100:~/itv-ghactivity
22/03/05 05:00:25 INFO YarnSchedulerBackend$YarnDriverEndpoint: Registered executor NettyRpcEndpointRef(spark-client://Executor) (172.31.15.174:60716) with ID 2, ResourceProfileId 0
22/03/05 05:00:25 INFO ExecutorMonitor: New executor 2 has registered (new total is 3)
22/03/05 05:00:25 INFO BlockManagerMasterEndpoint: Registering block manager ip-172-31-15-174.us-east-2.compute.internal:36885 with 4.8 GiB RAM, BlockManagerId(2, ip-172-31-15-174.us-east-2.compute.internal, 36885, None)
22/03/05 05:00:25 INFO TaskSetManager: Starting task 14.0 in stage 1.0 (TID 33) (ip-172-31-15-174.us-east-2.compute.internal, executor 2, partition 14, RACK_LOCAL, 5397 bytes) taskResourceAssignments Map()
22/03/05 05:00:25 INFO TaskSetManager: Starting task 15.0 in stage 1.0 (TID 34) (ip-172-31-15-174.us-east-2.compute.internal, executor 2, partition 15, RACK_LOCAL, 5396 bytes) taskResourceAssignments Map()
22/03/05 05:00:26 INFO BlockManagerInfo: Added broadcast_3_piece0 in memory on ip-172-31-15-174.us-east-2.compute.internal:36885 (size: 98.3 KiB, free: 4.8 GiB)
22/03/05 05:00:28 INFO BlockManagerInfo: Added broadcast_2_piece0 in memory on ip-172-31-15-174.us-east-2.compute.internal:36885 (size: 36.1 KiB, free: 4.8 GiB)
22/03/05 05:00:34 INFO TaskSetManager: Finished task 12.0 in stage 1.0 (TID 31) in 37129 ms on ip-172-31-5-189.us-east-2.compute.internal (executor 1) (11/16)
22/03/05 05:00:34 INFO TaskSetManager: Finished task 11.0 in stage 1.0 (TID 30) in 39382 ms on ip-172-31-5-189.us-east-2.compute.internal (executor 1) (12/16)
22/03/05 05:00:36 INFO TaskSetManager: Finished task 10.0 in stage 1.0 (TID 29) in 41784 ms on ip-172-31-5-189.us-east-2.compute.internal (executor 1) (13/16)
22/03/05 05:00:37 INFO TaskSetManager: Finished task 13.0 in stage 1.0 (TID 32) in 32320 ms on ip-172-31-5-189.us-east-2.compute.internal (executor 1) (14/16)
22/03/05 05:00:57 INFO TaskSetManager: Finished task 15.0 in stage 1.0 (TID 34) in 31502 ms on ip-172-31-15-174.us-east-2.compute.internal (executor 2) (15/16)
22/03/05 05:00:57 INFO TaskSetManager: Finished task 14.0 in stage 1.0 (TID 33) in 31870 ms on ip-172-31-15-174.us-east-2.compute.internal (executor 2) (16/16)
22/03/05 05:00:57 INFO YarnScheduler: Removed TaskSet 1.0, whose tasks have all completed, from pool
22/03/05 05:00:57 INFO DAGScheduler: ResultStage 1 (save at NativeMethodAccessorImpl.java:0) finished in 213.936 s
22/03/05 05:00:57 INFO DAGScheduler: Job 1 is finished. Cancelling potential speculative or zombie tasks for this job
22/03/05 05:00:57 INFO YarnScheduler: Killing all running tasks in stage 1: Stage finished
22/03/05 05:00:57 INFO DAGScheduler: Job 1 finished: save at NativeMethodAccessorImpl.java:0, took 213.944227 s
22/03/05 05:00:57 INFO MultipartUploadOutputStream: close closed:false s3://indu-github-emr/prod/raw/ghactivity/_SUCCESS
22/03/05 05:00:58 INFO FileFormatWriter: Write Job 06b9aefb-0cc9-4832-84bb-39064814fd3e committed.
22/03/05 05:00:58 INFO FileFormatWriter: Finished processing stats for write job 06b9aefb-0cc9-4832-84bb-39064814fd3e.
22/03/05 05:00:58 INFO SparkContext: Invoking stop() from shutdown hook
22/03/05 05:00:58 INFO AbstractConnector: Stopped Spark@7eb46217(HTTP/1.1, (http://1.1)}{0.0.0.0:4040}
22/03/05 05:00:58 INFO SparkUI: Stopped Spark web UI at http://ip-172-31-2-100.us-east-2.compute.internal:4040
22/03/05 05:00:58 INFO YarnClientSchedulerBackend: Interrupting monitor thread
22/03/05 05:00:58 INFO YarnClientSchedulerBackend: Shutting down all executors
22/03/05 05:00:58 INFO YarnSchedulerBackend$YarnDriverEndpoint: Asking each executor to shut down
22/03/05 05:00:58 INFO YarnClientSchedulerBackend: YARN client scheduler backend Stopped
22/03/05 05:00:58 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
22/03/05 05:00:58 INFO MemoryStore: MemoryStore cleared
22/03/05 05:00:58 INFO BlockManager: BlockManager stopped
22/03/05 05:00:58 INFO BlockManagerMaster: BlockManagerMaster stopped
22/03/05 05:00:58 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
22/03/05 05:00:58 INFO SparkContext: Successfully stopped SparkContext
22/03/05 05:00:58 INFO ShutdownHookManager: Shutdown hook called
22/03/05 05:00:58 INFO ShutdownHookManager: Deleting directory /mnt/tmp/spark-57e56f23-e16f-47fc-af13-5cad667236f5
22/03/05 05:00:58 INFO ShutdownHookManager: Deleting directory /mnt/tmp/spark-d0fe5f04-0be7-42aa-ad46-c955569f4f58
22/03/05 05:00:58 INFO ShutdownHookManager: Deleting directory /mnt/tmp/spark-d0fe5f04-0be7-42aa-ad46-c955569f4f58/pyspark-49394fd7-c4da-4a8c-a520-5cab1d7ac87
[hadoop@ip-172-31-2-100 itv-ghactivity]$
```

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

