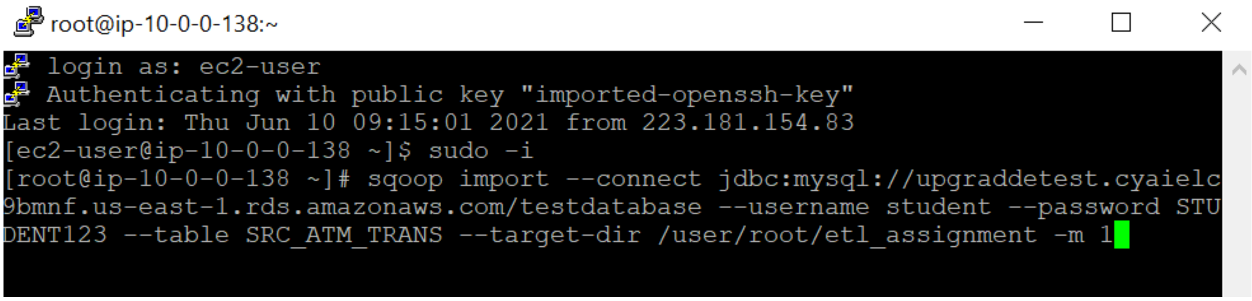
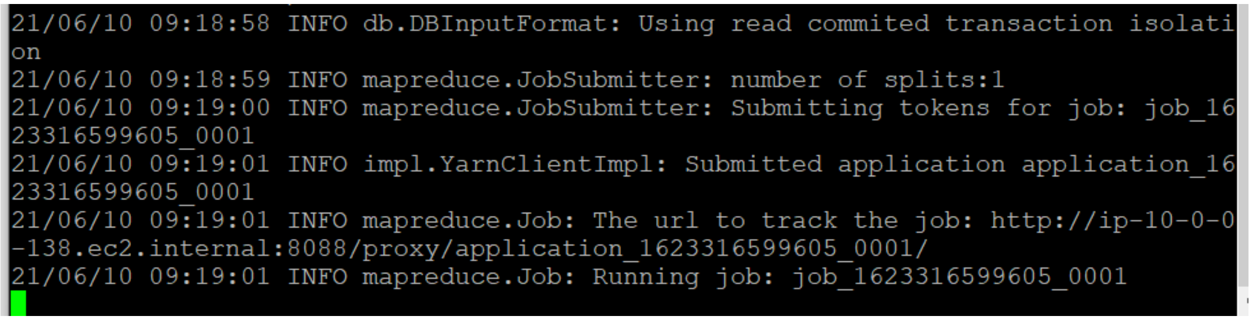
Data Ingestion from the RDS to HDFS using Sqoop

Sqoop Import command used for importing table from RDS to HDFS:

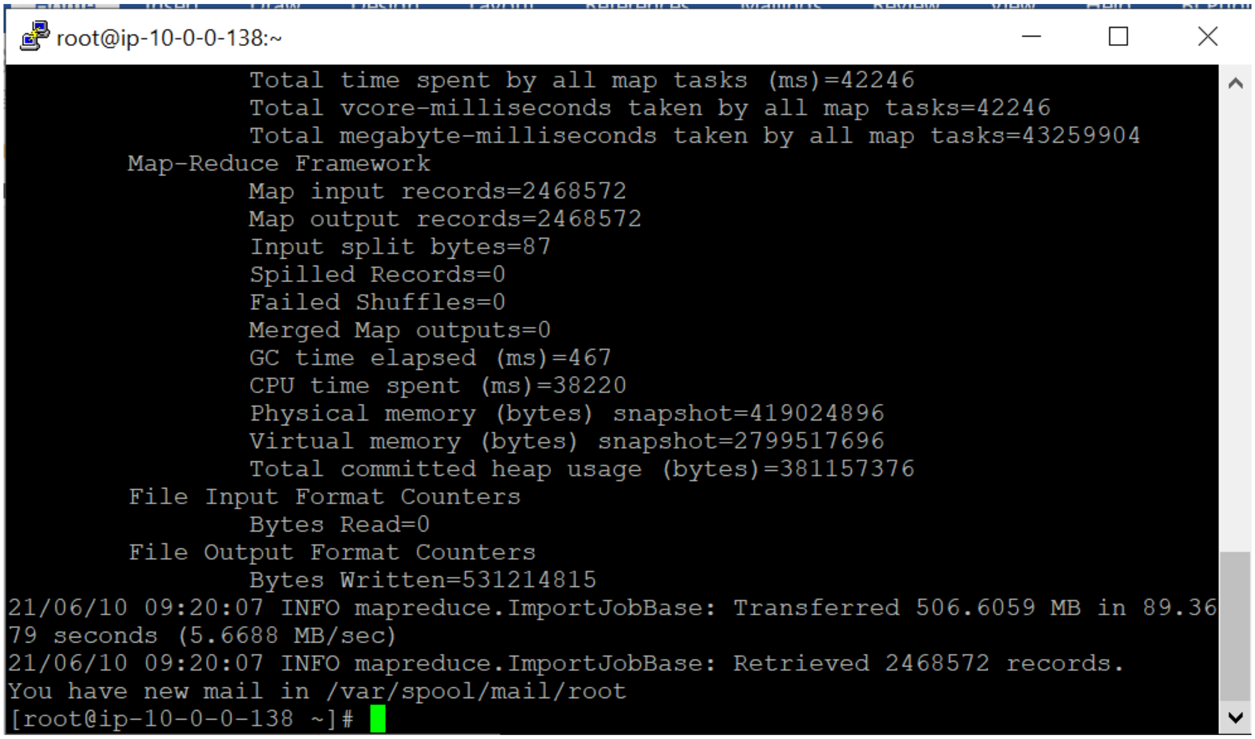
sqoop import --connect jdbc:mysql://upgraddetest.cyaielc9bmnf.us-east-1.rds.amazonaws.com/testdatabase --username student --password STUDENT123 --table SRC_ATM_TRANS --target-dir /user/root/etl_assignment -m 1



```
 login as: ec2-user
 Authenticating with public key "imported-openssh-key"
Last login: Thu Jun 10 09:15:01 2021 from 223.181.154.83
[ec2-user@ip-10-0-0-138 ~]$ sudo -i
[root@ip-10-0-0-138 ~]# sqoop import --connect jdbc:mysql://upgraddetest.cyaielc
9bmnf.us-east-1.rds.amazonaws.com/testdatabase --username student --password STU
DENT123 --table SRC_ATM_TRANS --target-dir /user/root/etl_assignment -m 1
```



```
21/06/10 09:18:58 INFO db.DBInputFormat: Using read commited transaction isolati
on
21/06/10 09:18:59 INFO mapreduce.JobSubmitter: number of splits:1
21/06/10 09:19:00 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_16
23316599605_0001
21/06/10 09:19:01 INFO impl.YarnClientImpl: Submitted application application_16
23316599605_0001
21/06/10 09:19:01 INFO mapreduce.Job: The url to track the job: http://ip-10-0-0
-138.ec2.internal:8088/proxy/application_1623316599605_0001/
21/06/10 09:19:01 INFO mapreduce.Job: Running job: job_1623316599605_0001
```

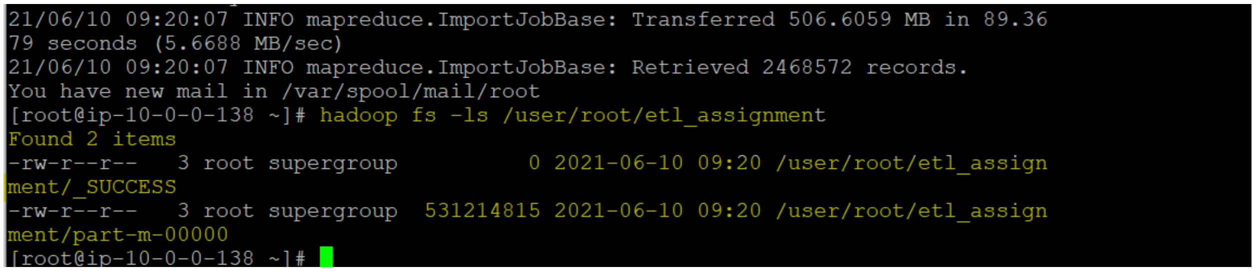


```
                Total time spent by all map tasks (ms)=42246
                Total vcore-milliseconds taken by all map tasks=42246
                Total megabyte-milliseconds taken by all map tasks=43259904
        Map-Reduce Framework
                Map input records=2468572
                Map output records=2468572
                Input split bytes=87
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=467
                CPU time spent (ms)=38220
                Physical memory (bytes) snapshot=419024896
                Virtual memory (bytes) snapshot=2799517696
                Total committed heap usage (bytes)=381157376
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=531214815
21/06/10 09:20:07 INFO mapreduce.ImportJobBase: Transferred 506.6059 MB in 89.36
79 seconds (5.6688 MB/sec)
21/06/10 09:20:07 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
You have new mail in /var/spool/mail/root
[root@ip-10-0-0-138 ~]#
```

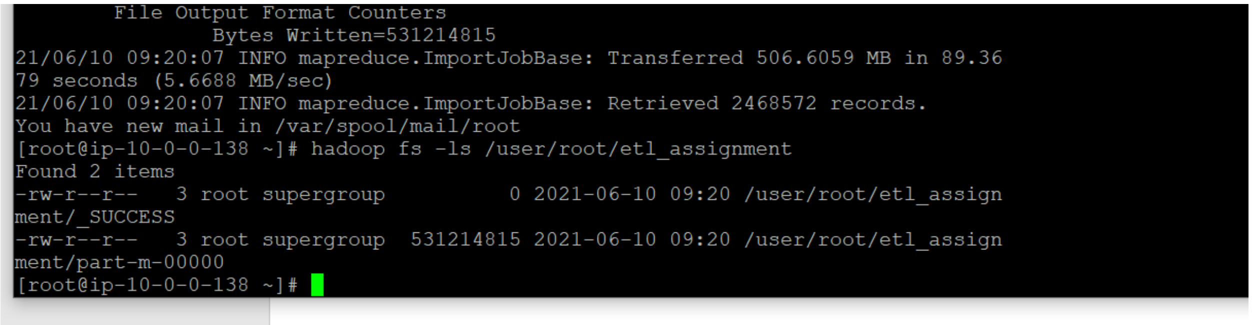
Command used to see the list of imported data in HDFS:

hadoop fs -ls /user/root/**etl_assignment**

Screenshot of the imported data:



```
21/06/10 09:20:07 INFO mapreduce.ImportJobBase: Transferred 506.6059 MB in 89.36
79 seconds (5.6688 MB/sec)
21/06/10 09:20:07 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
You have new mail in /var/spool/mail/root
[root@ip-10-0-0-138 ~]# hadoop fs -ls /user/root/etl_assignment
Found 2 items
-rw-r--r--   3 root supergroup          0 2021-06-10 09:20 /user/root/etl_assign
ment/_SUCCESS
-rw-r--r--   3 root supergroup  531214815 2021-06-10 09:20 /user/root/etl_assign
ment/part-m-00000
[root@ip-10-0-0-138 ~]#
```



```
        File Output Format Counters
                Bytes Written=531214815
21/06/10 09:20:07 INFO mapreduce.ImportJobBase: Transferred 506.6059 MB in 89.36
79 seconds (5.6688 MB/sec)
21/06/10 09:20:07 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
You have new mail in /var/spool/mail/root
[root@ip-10-0-0-138 ~]# hadoop fs -ls /user/root/etl_assignment
Found 2 items
-rw-r--r--   3 root supergroup          0 2021-06-10 09:20 /user/root/etl_assign
ment/_SUCCESS
-rw-r--r--   3 root supergroup  531214815 2021-06-10 09:20 /user/root/etl_assign
ment/part-m-00000
[root@ip-10-0-0-138 ~]#
```

Explanation:

1) Sqoop command imports RDS data table SRC_ATM_TRANS to a target directory located at root with a name “etl_assignment”
2) Map reduce job will be invoked
3) After the job successfully completes check the files after login as root user
4) Using command hadoop fs -ls /user/root/etl_assignment which is used to list all files present under etl_assignment
5) _SUCCESS, it says data from RDS is loaded sucessfully and there is a part file as we ran this m 1