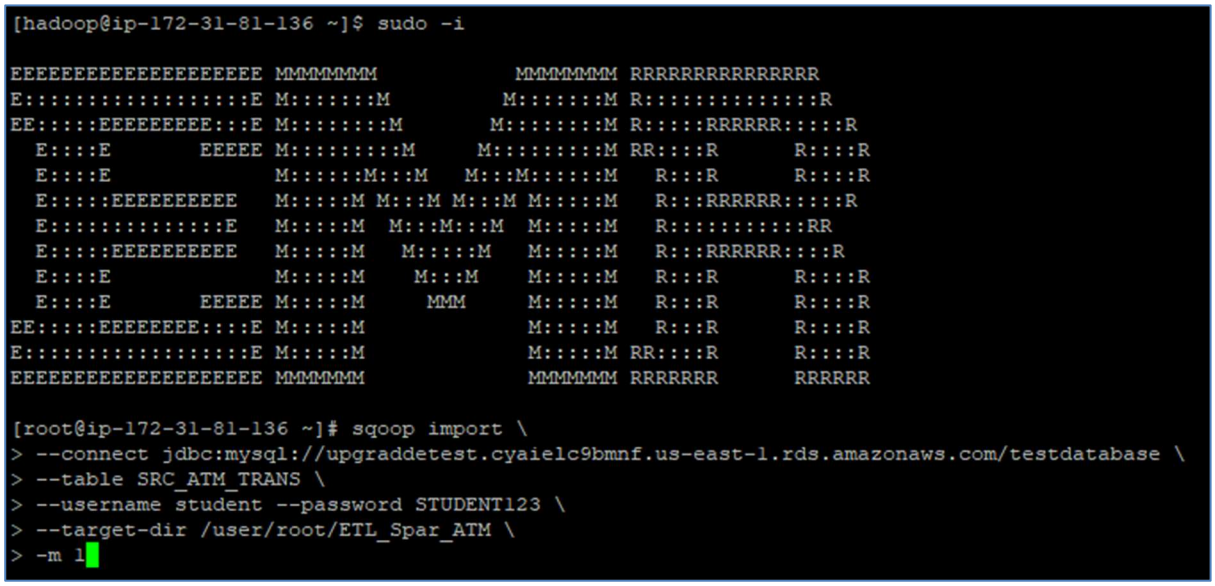
Data Ingestion from the RDS to HDFS using Sqoop

Before running Sqoop import command we have installed & Configure MySQL Connector.

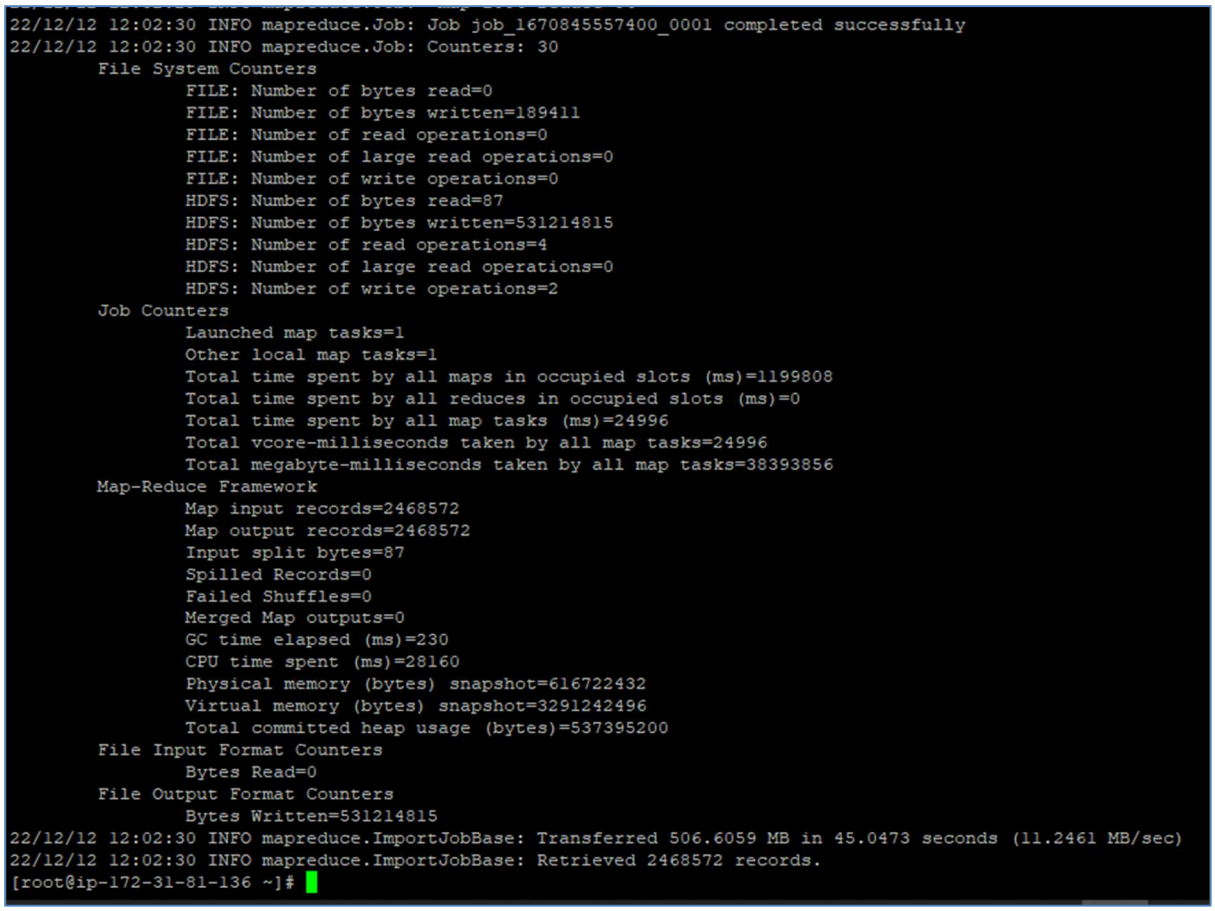
Sqoop Import command used for importing table from RDS to HDFS:

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

Screenshot of Sqoop import command Execution:



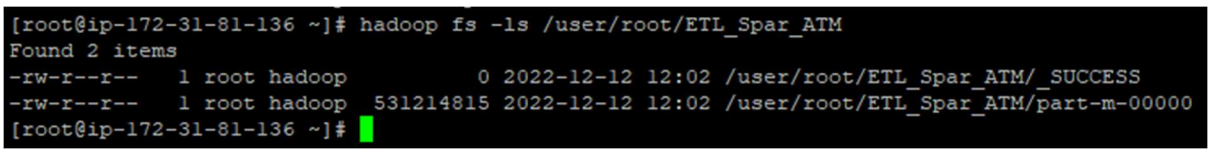
Screenshot of success of above import command:



In above screenshot we can see that total 246572 records have been imported.

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

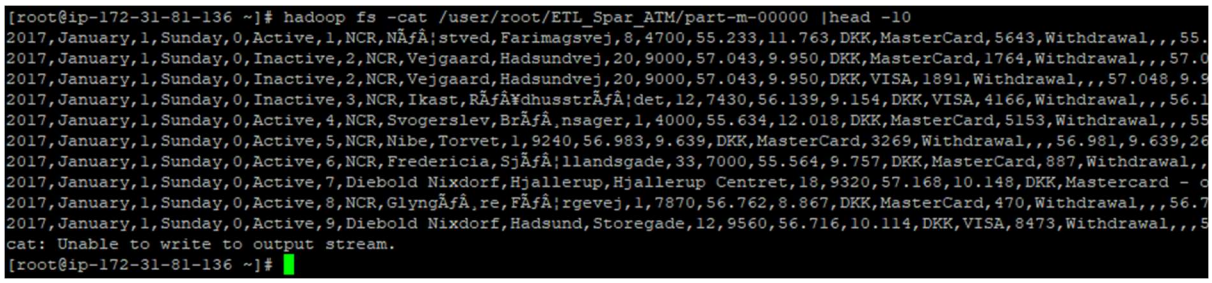
hadoop fs -ls /user/root/ETL_Spar_ATM



In the screenshot above we can see two items:
- The first file is the success file, indicating that the MapReduce job was successful.
- The second file ‘part-m-00000’ is the one that we imported. Since we used only one mapper in my import command, thus the data is in a single file.

Screenshot of the imported data:

hadoop fs -cat /user/root/ETL_Spar_ATM/part-m-00000 |head -10



With above command we are checking 10 rows of imported data.

Now below location and file we will use in PySpark for data transformation:
/user/root/ETL_Spar_ATM/part-m-00000