Data Ingestion from the RDS to HDFS using Sqoop

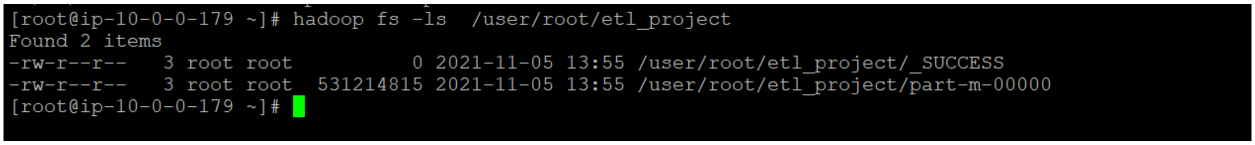
Sqoop Import command used for importing table from RDS to HDFS:

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

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

```
hadoop fs -ls  /user/root/etl_project
```

Screenshot of the imported data:

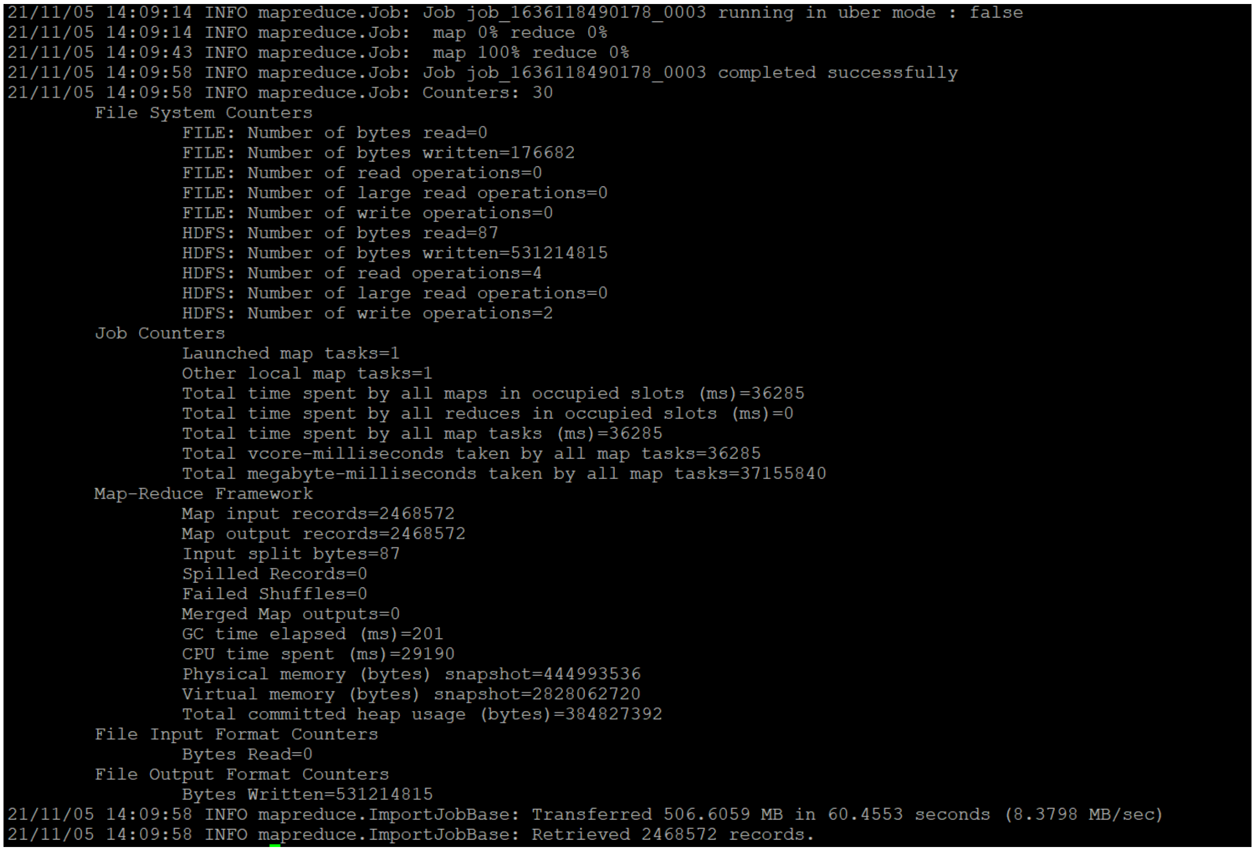


```
[root@ip-10-0-0-179 ~]# hadoop fs -ls  /user/root/etl_project
Found 2 items
-rw-r--r--   3 root root          0 2021-11-05 13:55 /user/root/etl_project/_SUCCESS
-rw-r--r--   3 root root  531214815 2021-11-05 13:55 /user/root/etl_project/part-m-00000
[root@ip-10-0-0-179 ~]#
```

Explanation and comments:

1. We logged into ec2 instance and switched to root user before starting off with importing of data from RDS.
2. We have imported data from given rds instance to hdfs using sqoop import command with 1 mapper job.
3. We mentioned the target directory in sqoop command as **/user/root/etl_project**. Once sqoop finishes importing the data it gives info on number of map jobs and reduce jobs, in this case reduce is always 0. And in the end shows total size of the table

transferred, number of rows it has and time taken to transfer.



```
21/11/05 14:09:14 INFO mapreduce.Job: Job job_1636118490178_0003 running in uber mode : false
21/11/05 14:09:14 INFO mapreduce.Job:  map 0% reduce 0%
21/11/05 14:09:43 INFO mapreduce.Job:  map 100% reduce 0%
21/11/05 14:09:58 INFO mapreduce.Job: Job job_1636118490178_0003 completed successfully
21/11/05 14:09:58 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=176682
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=87
                HDFS: Number of bytes written=531214815
                HDFS: Number of read operations=4
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Other local map tasks=1
                Total time spent by all maps in occupied slots (ms)=36285
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=36285
                Total vcore-milliseconds taken by all map tasks=36285
                Total megabyte-milliseconds taken by all map tasks=37155840
        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)=201
                CPU time spent (ms)=29190
                Physical memory (bytes) snapshot=444993536
                Virtual memory (bytes) snapshot=2828062720
                Total committed heap usage (bytes)=384827392
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=531214815
21/11/05 14:09:58 INFO mapreduce.ImportJobBase: Transferred 506.6059 MB in 60.4553 seconds (8.3798 MB/sec)
21/11/05 14:09:58 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
```

4. Once sqoop import is done we verified the data is present in hdfs by typing -> hadoop fs -ls /user/root/etl_project. This shows two outputs -> _SUCEESS indicating sqoop import was successful and next is part-m-00000 as we used only 1 mapper job there is only 1 part.