Data Ingestion from the RDS to HDFS using Sqoop

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

The command to import data from RDS to HDFS is:

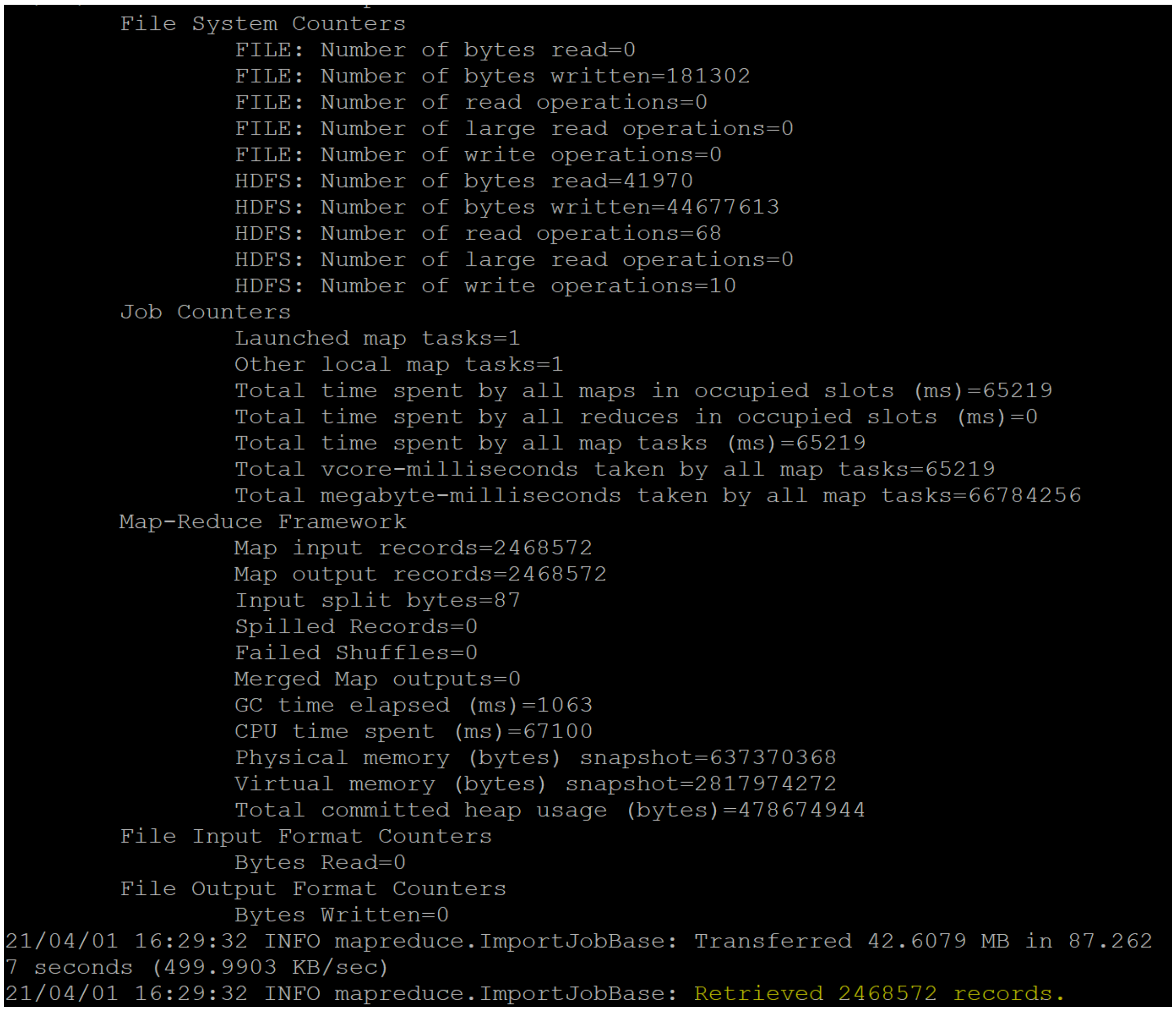
```
sqoop import –connect <RDS connection string> \
--username <RDS Username> --password <RDS Password> \
--table <Table name in RDS> \
--null-string '\\N' --null-non-string '\\N'
--target-dir <path in HDFS where the data is imported to> \
-m <no. of mapper jobs> --as-parquetfile
```

Here,
RDS connect string: “jdbc:mysql://upgraddetest.cyaielc9bmnf.us-east-1.rds.amazonaws.com/testdatabase”
Username: “student”
Password: “STUDENT123”
Table name: “SRC_ATM_TRANS” and
Target dir: “/user/root/atm_trans”

We are saving this as a “parquet” file and running only 1 mapper job. So the sqoop import command will be as given below:

```
sqoop import \
--connect jdbc:mysql://upgraddetest.cyaielc9bmnf.us-east-1.rds.amazonaws.com/testdatabase \
--table SRC_ATM_TRANS \
--username student --password STUDENT123 \
--null-string ‘\\N’ –null-non-string ‘\\N’
--target-dir /user/root/atm_trans \
-m 1 –as-parquetfile
```

© Copyright 2020. upGrad Education Pvt. Ltd. All rights reserved



```
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=181302
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=41970
                HDFS: Number of bytes written=44677613
                HDFS: Number of read operations=68
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=10
        Job Counters
                Launched map tasks=1
                Other local map tasks=1
                Total time spent by all maps in occupied slots (ms)=65219
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=65219
                Total vcore-milliseconds taken by all map tasks=65219
                Total megabyte-milliseconds taken by all map tasks=66784256
        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)=1063
                CPU time spent (ms)=67100
                Physical memory (bytes) snapshot=637370368
                Virtual memory (bytes) snapshot=2817974272
                Total committed heap usage (bytes)=478674944
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
21/04/01 16:29:32 INFO mapreduce.ImportJobBase: Transferred 42.6079 MB in 87.262
7 seconds (499.9903 KB/sec)
21/04/01 16:29:32 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
```

We see that all the records (2468572) are successfully imported to HDFS.

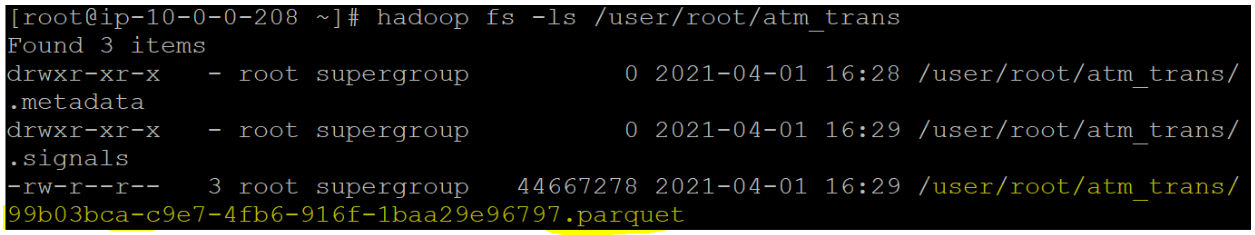
Command used to see the list of imported data in HDFS:

Below is the command to see the files present in the directory atm_trans, i.e., the HDFS location to which we imported the data:

hadoop fs -ls /user/root/atm_trans

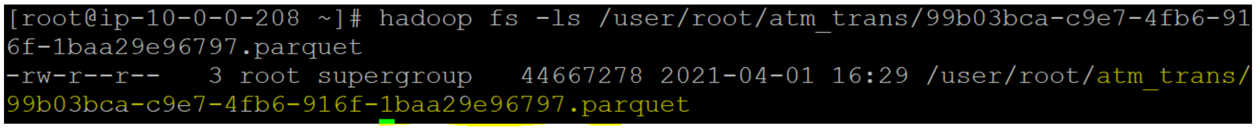
Where “/user/root/atm_trans” is the location where the data is imported in to the HDFS and “99b03bca-c9e7-4fb6-916f-1baa29e96797.parquet” is the record. So, we pass the full path in the “hadoop fs -ls” command. The data imported looks as shown below.

© Copyright 2020. upGrad Education Pvt. Ltd. All rights reserved



```
[root@ip-10-0-0-208 ~]# hadoop fs -ls /user/root/atm_trans
Found 3 items
drwxr-xr-x   - root supergroup          0 2021-04-01 16:28 /user/root/atm_trans/
.metadata
drwxr-xr-x   - root supergroup          0 2021-04-01 16:29 /user/root/atm_trans/
.signals
-rw-r--r--   3 root supergroup   44667278 2021-04-01 16:29 /user/root/atm_trans/
99b03bca-c9e7-4fb6-916f-1baa29e96797.parquet
```

Screenshot of the imported data:
Let us check the parquet file that was created during sqoop import and see if the required file is present in it.



```
[root@ip-10-0-0-208 ~]# hadoop fs -ls /user/root/atm_trans/99b03bca-c9e7-4fb6-91
6f-1baa29e96797.parquet
-rw-r--r--   3 root supergroup   44667278 2021-04-01 16:29 /user/root/atm_trans/
99b03bca-c9e7-4fb6-916f-1baa29e96797.parquet
```

© Copyright 2020. upGrad Education Pvt. Ltd. All rights reserved