

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://database-1.cnxebh8m2ub8.us-east-1.rds.amazonaws.com:3306/etl_data --table atm_data --columns "year,month,day,weekday,hour,atm_status,atm_id,atm_manufacturer,atm_location,atm_streetname,atm_street_number,atm_zipcode,atm_lat,atm_lon,currency,card_type,service,message_code,message_text,weather_lat,weather_lon,weather_city_id,weather_city_name,temp,pressure,humidity,wind_speed,wind_deg,rain_3h,clouds_all,weather_id,weather_main,weather_description" --username admin --target-dir /user/root/atm_dataset -m 1 -P
```

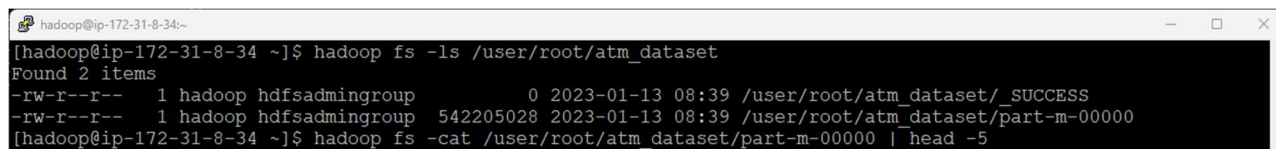
Line wise Comment

2. **--connect**: Connecting to RDS using '**jdbc**' connector to '**mysql**' server. '**database-1.cnxebh8m2ub8.us-east-1.rds.amazonaws.com:3306**' is the RDS public dns and port number. '**etl_data**' is the name of the Database.
3. **--table**: The required data is stored on the table '**atm_data**'.
4. **--column**: This line is for arranging the columns.
If this line is not used the columns will be arrange in a sorted manner.
5. **--username**: '**admin**' is the username used for logging into the RDS.
6. **--target-dir**: '**/user/root/atm_dataset**' is selected to be the location in hdfs where the importing data must be stored.
7. **-m 1**: The number of mappers to be used here is only 1.
8. **-P**: The job will ask you the password when connecting to RDS server.

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

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

Screenshot of the imported data:



```
hadoop@ip-172-31-8-34:~$ hadoop fs -ls /user/root/atm_dataset
Found 2 items
-rw-r--r-- 1 hadoop hdfsadmin group 0 2023-01-13 08:39 /user/root/atm_dataset/_SUCCESS
-rw-r--r-- 1 hadoop hdfsadmin group 542205028 2023-01-13 08:39 /user/root/atm_dataset/part-m-00000
hadoop@ip-172-31-8-34:~$ hadoop fs -cat /user/root/atm_dataset/part-m-00000 | head -5
```