

## 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.cyaieic9bmnf.us-east-1.rds.amazonaws.com/testdatabase \  
--username student --password STUDENT123 --table SRC_ATM_TRANS \  
--m 1 --as-textfile \  
--target-dir /user/root/etl_assignment \  
--null-string '\N' --null-non-string '\N'
```

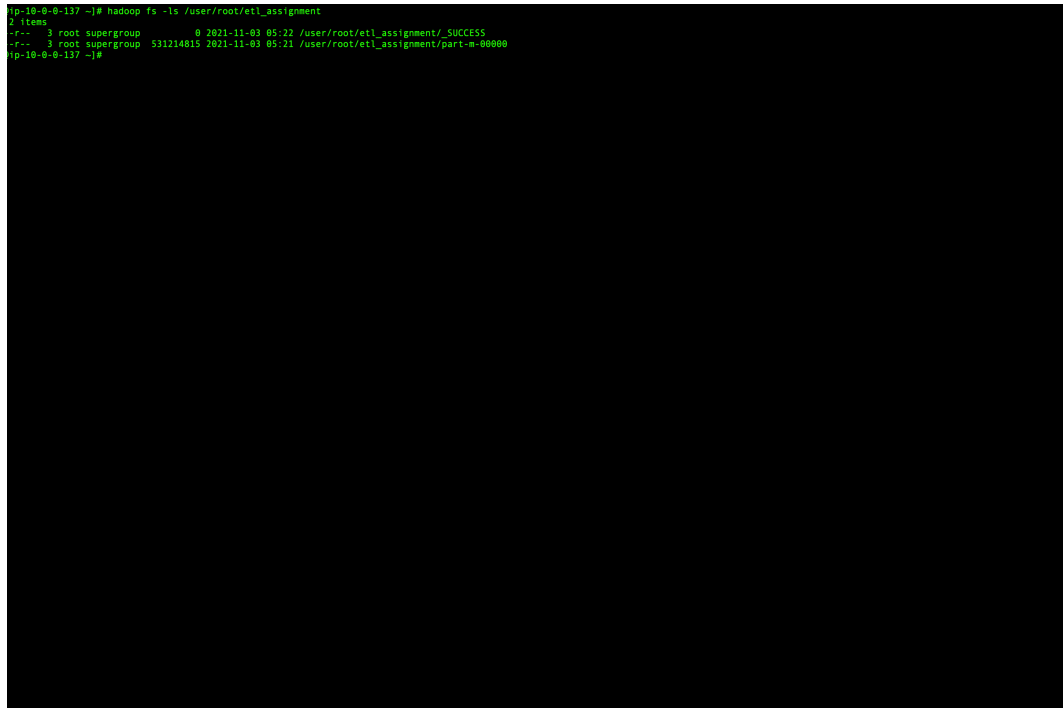
```
ip-10-0-0-137 ~# sqoop import \  
--connect jdbc:mysql://upgraddetest.cyaieic9bmnf.us-east-1.rds.amazonaws.com/testdatabase \  
--username student --password STUDENT123 --table SRC_ATM_TRANS \  
--m 1 --as-textfile \  
--target-dir /user/root/etl_assignment \  
--null-string '\N' --null-non-string '\N'  
05:20:49 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.15.1  
05:20:49 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.  
05:20:49 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.  
05:20:49 INFO tool.CodeGenTool: Beginning code generation  
05:20:50 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM 'SRC_ATM_TRANS' AS t LIMIT 1  
05:20:50 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM 'SRC_ATM_TRANS' AS t LIMIT 1  
05:20:50 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /opt/cloudera/parcels/CDH/lib/hadoop-mapreduce  
/tmp/sqoop-root/compile/3648c7a32a7c1adac69625bdb0488818/SRC_ATM_TRANS.java uses or overrides a deprecated API.  
Recompile with -Xlint:deprecation for details.  
05:20:55 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-root/compile/3648c7a32a7c1adac69625bdb0488818/SRC_ATM_TRANS.jar  
05:20:55 WARN manager.MySQLManager: It looks like you are importing from mysql.  
05:20:55 WARN manager.MySQLManager: This transfer can be faster! Use the --direct  
05:20:55 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.  
05:20:55 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)  
05:20:55 INFO mapreduce.ImportJobBase: Beginning import of SRC_ATM_TRANS  
05:20:55 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar  
05:20:56 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps  
05:20:56 INFO client.RMProxy: Connecting to ResourceManager at ip-10-0-0-137.ec2.internal/10.0.0.137:8032  
05:21:05 INFO db.DBInputFormat: Using read committed transaction isolation  
05:21:05 INFO mapreduce.JobSubmitter: number of splits:1  
05:21:05 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1635916520014_0001  
05:21:06 INFO Impl.YarnClientImpl: Submitted application application_1635916520014_0001  
05:21:06 INFO mapreduce.Job: The url to track the job: http://ip-10-0-0-137.ec2.internal:8088/proxy/application_1635916520014_0001/  
05:21:06 INFO mapreduce.Job: Running job: job_1635916520014_0001
```

Running the Sqoop Command

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

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

**Screenshot of the imported data:**

A terminal window screenshot showing the execution of the command 'hadoop fs -ls /user/root/etl\_assignment'. The output lists two items: a directory named '\_SUCCESS' and a file named 'part-m-00000'. The directory listing includes permissions, owner, group, size, modification time, and the full path. The file listing includes permissions, owner, group, size, modification time, and the full path.

```
mp-18-0-0-137 ~]# hadoop fs -ls /user/root/etl_assignment
-rw-r--r-- 3 root supergroup          0 2021-11-03 05:22 /user/root/etl_assignment/_SUCCESS
-rw-r--r-- 3 root supergroup 531214815 2021-11-03 05:21 /user/root/etl_assignment/part-m-00000
mp-18-0-0-137 ~]#
```

Data imported to HDFS post Sqoop command

**Steps Taken:**

1. Login to the ec2 instance , run the sudo command
2. Execute the Sqoop command to import the data in the RDS table SRC\_ATM\_TRANS to the target directory user/root/etl\_assignment
3. Once the job completes navigate to the target directory set to find the imported data using the “hadoop fs -ls user/root/etl\_assignment” command
4. Since we see the \_SUCCESS it means the import has gone through successfully