Data Ingestion from the RDS to HDFS using Sqoop

1. Create an EMR Cluster with
    a. Hadoop, Zeppelin, Sqoop, Livy, Spark, Jupiter
    b. 1 single cluster
2. Set the respective Security Group

Connect to EMR Cluster using SSH	`ssh -i RHEL_new_1.pem hadoop@ec2-44-203-61-84.compute1.amazonaws.com`
Download the MySQL Connector	`wget https://de-mysql-connector.s3.amazonaws.com/mysqlconnector-java-8.0.25.tar.gz`
Unzip connecto r	`tar -xvf mysql-connector-java8.0.25.tar.gz`
Change to directory	`cd mysql-connector-java-8.0.25/`
Copy the jar file to Sqoop directory	`sudo cp mysql-connector-java-8.0.25.jar /usr/lib/sqoop/lib/`

3. **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 --username student --password STUDENT123 --table SRC_ATM_TRANS --target-dir /home/data -m 1
```

```
[hadoop@ip-172-31-6-72 mysql-connector-java-8.0.25]$ sqoop import --connect jdbc:mysql://upgraddetest.cyaielc9bmnf.us-east-1.rds.amazonaws.com/testdatabase --username student --password STUDENT123 --table SRC_ATM_TRANS --target-dir /home/data -m 1
```

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

```
22/09/25 18:30:54 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=189779
                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)=1117728
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=23286
                Total vcore-milliseconds taken by all map tasks=23286
                Total megabyte-milliseconds taken by all map tasks=35767296
        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)=244
                CPU time spent (ms)=26040
                Physical memory (bytes) snapshot=627769344
                Virtual memory (bytes) snapshot=3292041216
                Total committed heap usage (bytes)=535298048
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=531214815
22/09/25 18:30:54 INFO mapreduce.ImportJobBase: Transferred 506.6059 MB in 43.7517 seconds (11.5791 MB/sec)
22/09/25 18:30:54 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
```

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

   ```
   hadoop fs –ls /home/data
   ```

```
[[hadoop@ip-172-31-6-72 mysql-connector-java-8.0.25]$ hadoop fs -ls /home/data
Found 2 items
-rw-r--r--   1 hadoop hadoop          0 2022-09-25 18:30 /home/data/_SUCCESS
-rw-r--r--   1 hadoop hadoop  531214815 2022-09-25 18:30 /home/data/part-m-00000
```

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

5. **Screenshot of the imported data:**

```
hadoop fs –cat /home/data/part–m–00000 |head
```

```
[hadoop@ip-172-31-6-72 mysql-connector-java-8.0.25]$ hadoop fs -cat /home/data/part-m-00000 |head
2017,January,1,Sunday,0,Active,1,NCR,NÃƒÂ¦stved,Farimagsvej,8,4700,55.233,11.763,DKK,MasterCard,5643,Withdrawal,,,55.230
,11.761,2616038,Naestved,281.150,1014,87,7,260,0.215,92,500,Rain,light rain
2017,January,1,Sunday,0,Inactive,2,NCR,Vejgaard,Hadsundvej,20,9000,57.043,9.950,DKK,MasterCard,1764,Withdrawal,,,57.048,
9.935,2616235,NÃƒÂ¸rresundby,280.640,1020,93,9,250,0.590,92,500,Rain,light rain
2017,January,1,Sunday,0,Inactive,2,NCR,Vejgaard,Hadsundvej,20,9000,57.043,9.950,DKK,VISA,1891,Withdrawal,,,57.048,9.935,
2616235,NÃƒÂ¸rresundby,280.640,1020,93,9,250,0.590,92,500,Rain,light rain
2017,January,1,Sunday,0,Inactive,3,NCR,Ikast,RÃƒÂ¥dhusstrÃƒÂ¦det,12,7430,56.139,9.154,DKK,VISA,4166,Withdrawal,,,56.139,
9.158,2619426,Ikast,281.150,1011,100,6,240,0.000,75,300,Drizzle,light intensity drizzle
2017,January,1,Sunday,0,Active,4,NCR,Svogerslev,BrÃƒÂ¸nsager,1,4000,55.634,12.018,DKK,MasterCard,5153,Withdrawal,,,55.64
2,12.080,2614481,Roskilde,280.610,1014,87,7,260,0.000,88,701,Mist,mist
2017,January,1,Sunday,0,Active,5,NCR,Nibe,Torvet,1,9240,56.983,9.639,DKK,MasterCard,3269,Withdrawal,,,56.981,9.639,26164
83,Nibe,280.640,1020,93,9,250,0.590,92,500,Rain,light rain
2017,January,1,Sunday,0,Active,6,NCR,Fredericia,SjÃƒÂ¦llandsgade,33,7000,55.564,9.757,DKK,MasterCard,887,Withdrawal,,,55
.566,9.753,2621951,Fredericia,281.150,1014,93,7,230,0.290,92,500,Rain,light rain
2017,January,1,Sunday,0,Active,7,Diebold Nixdorf,Hjallerup,Hjallerup Centret,18,9320,57.168,10.148,DKK,Mastercard - on-u
s,4626,Withdrawal,,,57.165,10.146,2620275,Hjallerup,280.640,1020,93,9,250,0.590,92,500,Rain,light rain
2017,January,1,Sunday,0,Active,8,NCR,GlyngÃƒÂ¸re,FÃƒÂ¦rgevej,1,7870,56.762,8.867,DKK,MasterCard,470,Withdrawal,,,56.793,
8.853,2615964,Nykobing Mors,281.150,1011,100,6,240,0.000,75,300,Drizzle,light intensity drizzle
2017,January,1,Sunday,0,Active,9,Diebold Nixdorf,Hadsund,Storegade,12,9560,56.716,10.114,DKK,VISA,8473,Withdrawal,,,56.7
15,10.117,2620952,Hadsund,280.640,1020,93,9,250,0.590,92,500,Rain,light rain
```

6. Move the data to the livy for using the data to the pyspark

```
hadoop fs –cp /home/data /user/livy
```

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