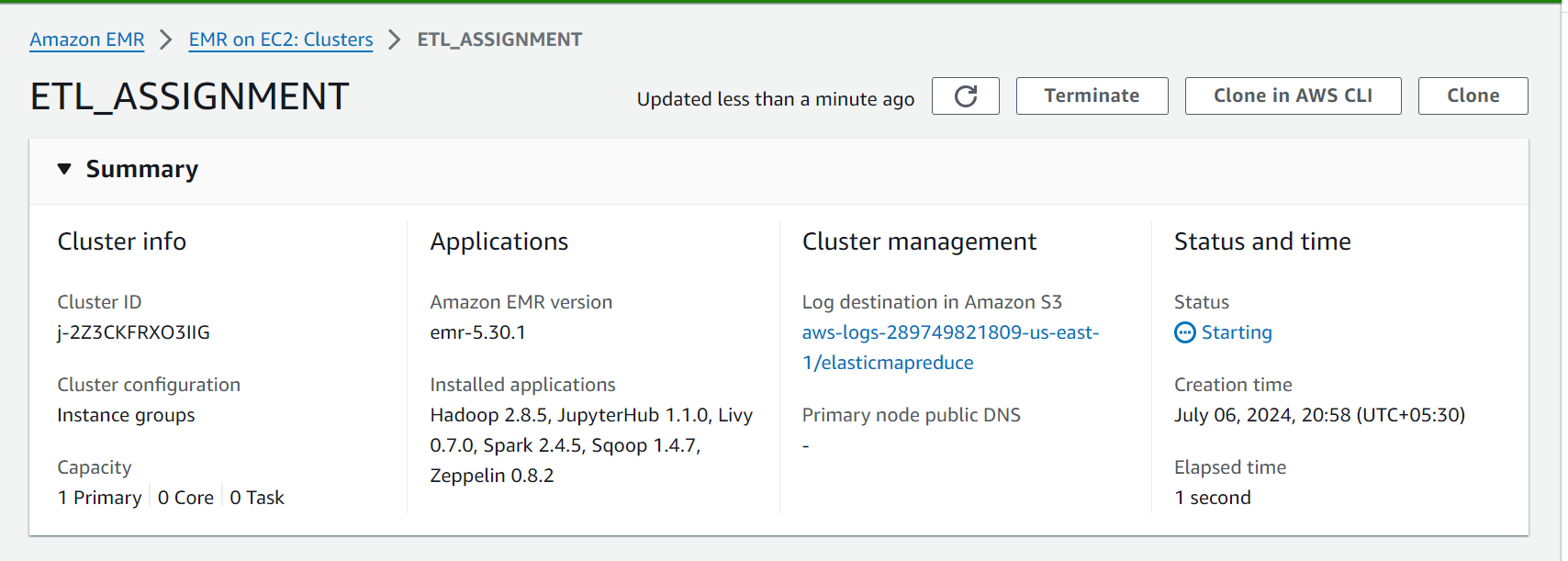
Apache Sqoop SQL to HADOOP

Sqoop used MapReduce to import and export the data, which provides parallel operation as well as fault tolerance.

EMR CLUSTER CREATION



SQL CONNECTOR INSTALLATION

wget <https://de-mysql-connector.s3.amazonaws.com/mysql-connector-java-8.0.25.tar.gz>

tar -xvf mysql-connector-java-8.0.25.tar.gz

cd mysql-connector-java-8.0.25/

sudo cp mysql-connector-java-8.0.25.jar /usr/lib/sqoop/lib/

mysql\_secure\_installation

mysql -u root -p

MariaDB [(none)]> GRANT ALL PRIVILEGES ON \*.\* TO 'root'@'%' identified by '1234' WITH GRANT OPTION;

Query OK, 0 rows affected (0.00 sec)

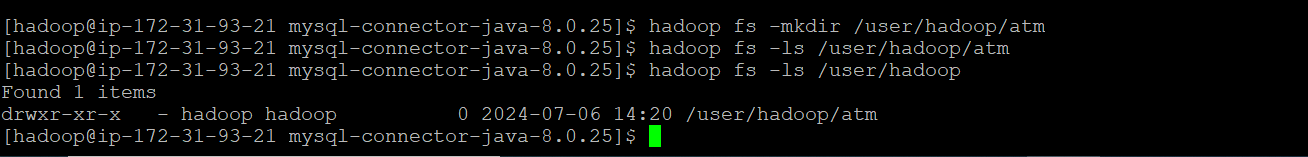
[hadoop@ip-172-31-93-21 mysql-connector-java-8.0.25]$ sudo service mysqld restart

Redirecting to /bin/systemctl restart mysqld.service

hadoop fs -mkdir /user/hadoop/atm

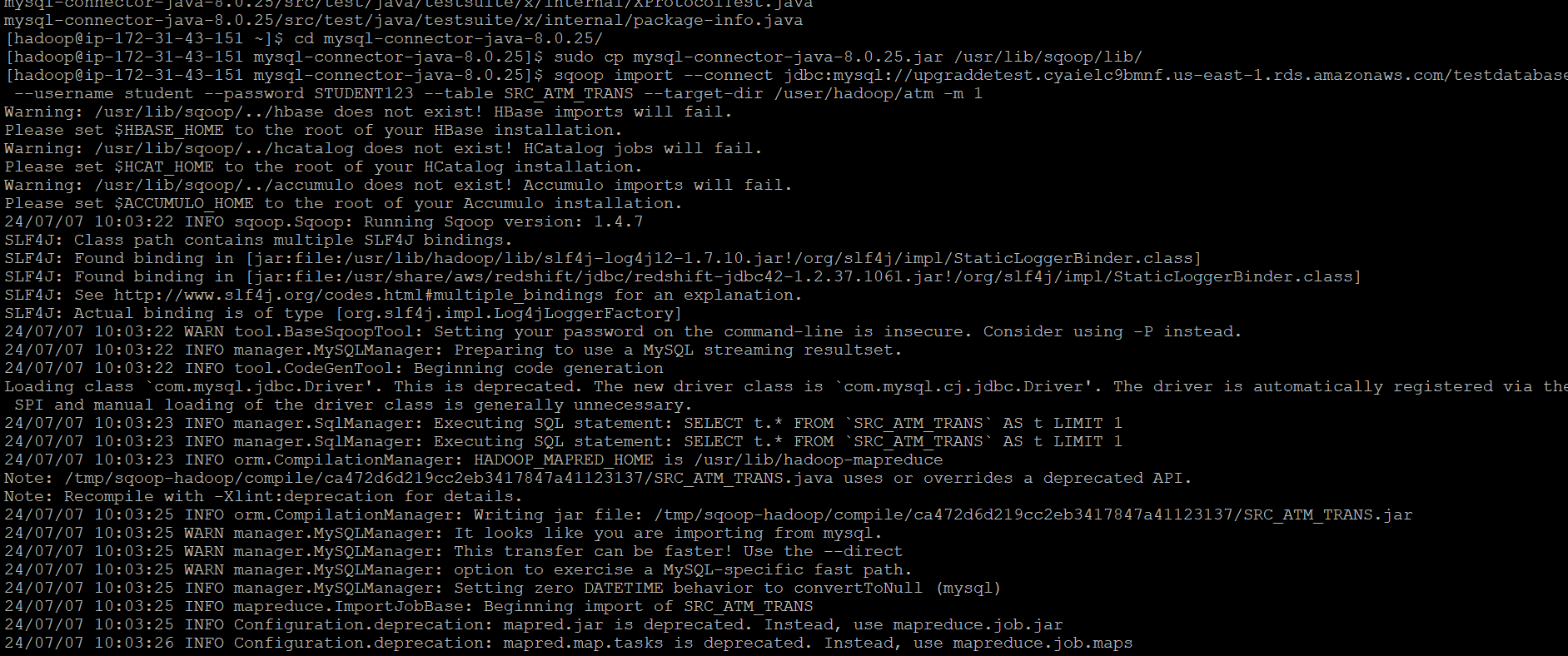
hadoop fs -ls /user/Hadoop

24/07/06 14:22:34 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7



**Note**: Before running the Sqoop import command, you need to ensure that the target directory does not already exist in HDFS. Otherwise, an error will be thrown that the target directory does not exist.

sqoop import --connect jdbc:mysql://upgraddetest.cyaielc9bmnf.us-east-1.rds.amazonaws.com/testdatabase --username student --password STUDENT123 --table SRC\_ATM\_TRANS --target-dir /user/hadoop/atm -m 1



A screenshot of a computer program

Description automatically generated

File Input Format Counters

Bytes Read=0

File Output Format Counters

Bytes Written=531214815

24/07/07 10:04:18 INFO mapreduce.ImportJobBase: Transferred 506.6059 MB in 51.9417 seconds (9.7534 MB/sec)

24/07/07 10:04:18 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.

[hadoop@ip-172-31-43-151 mysql-connector-java-8.0.25]$

**Data Ingestion with Sqoop ● Please check the number of records that are imported after the Sqoop Job. Count of Records – 2468572**

[hadoop@ip-172-31-93-21 mysql-connector-java-8.0.25]$ hadoop fs -cat /user/hadoop/atm/part-m-\* | head -n 10

A screen shot of a computer

Description automatically generated

VERIFY THE NUMBER OF ROWS AFTER EXPORTING DATA FROM HADOOP TO SPARK DATAFRAME

Data Processing with Spark

● Check count after importing data into a data frame.

A screenshot of a computer

Description automatically generated

● Check count for the Location Dimension

A screenshot of a computer

Description automatically generated

● Check count for the Card Type Dimension

A screenshot of a computer

Description automatically generated

● Check count for the ATM Dimension

A screenshot of a computer program

Description automatically generated

● Check count for the Date Dimension

A screenshot of a computer

Description automatically generated

● Check count for the all the Stages in the creation of Transaction Fact table

