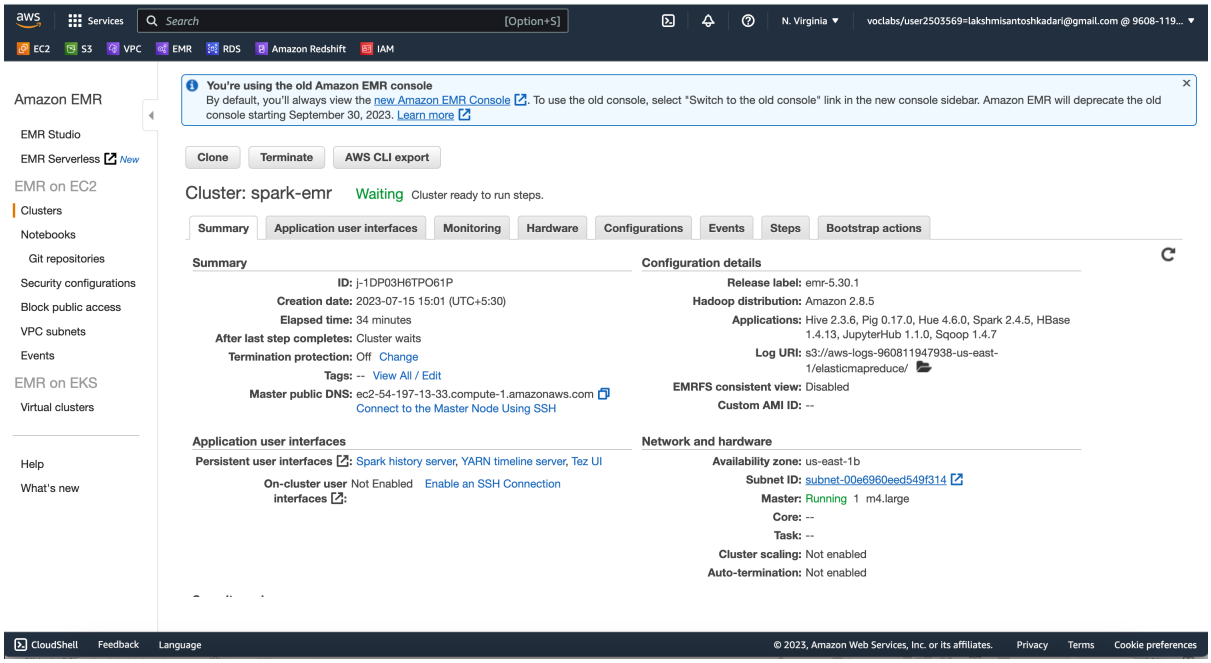
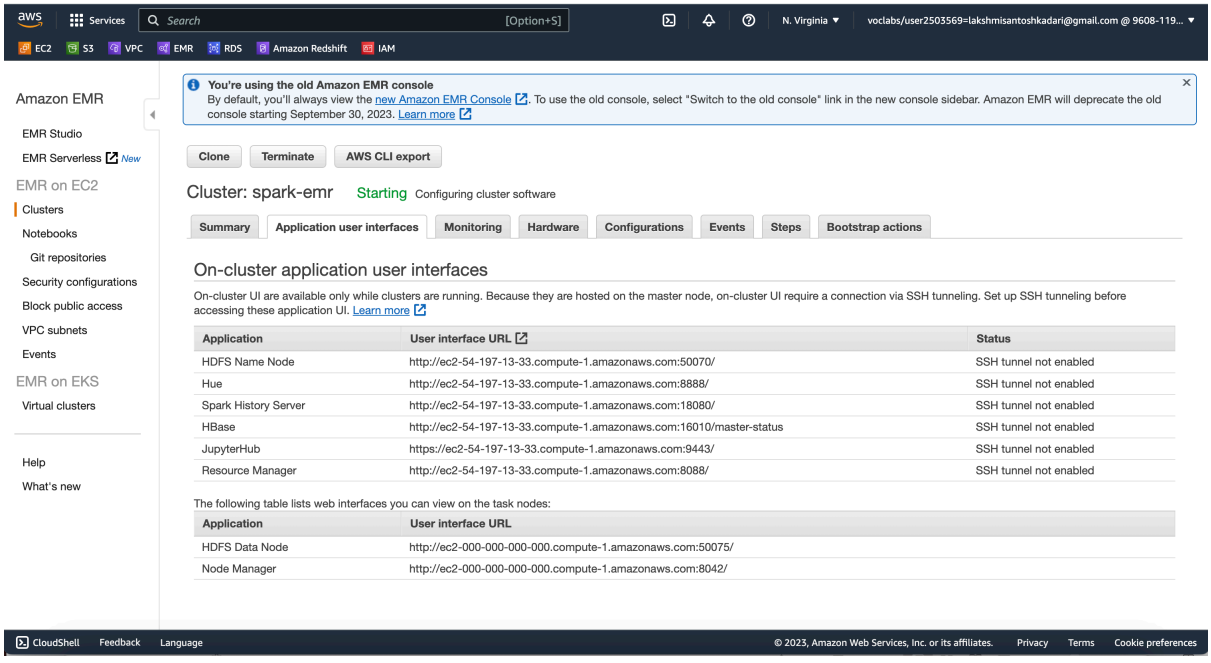
Data Ingestion

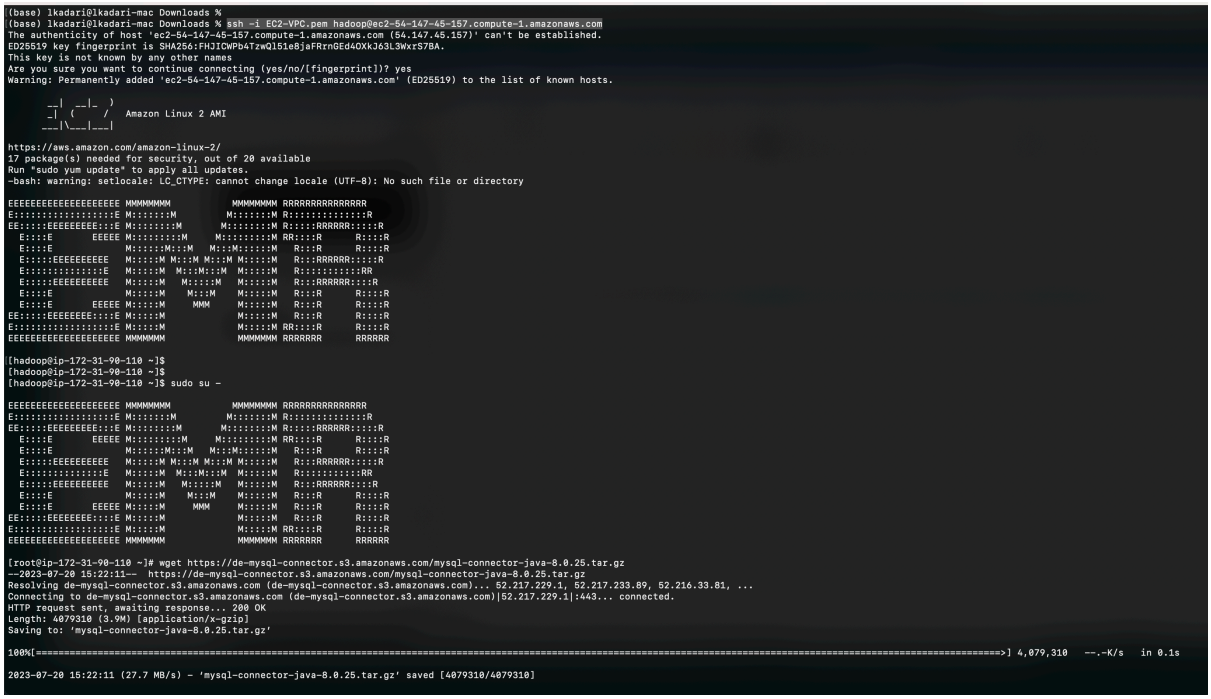
Step 1: Created an EMR cluster as shown in the screenshots.





Step 2: Logging in to the cluster using below command.

ssh -i EC2-VPC.pem hadoop@ec2-54-147-45-157.compute-1.amazonaws.com

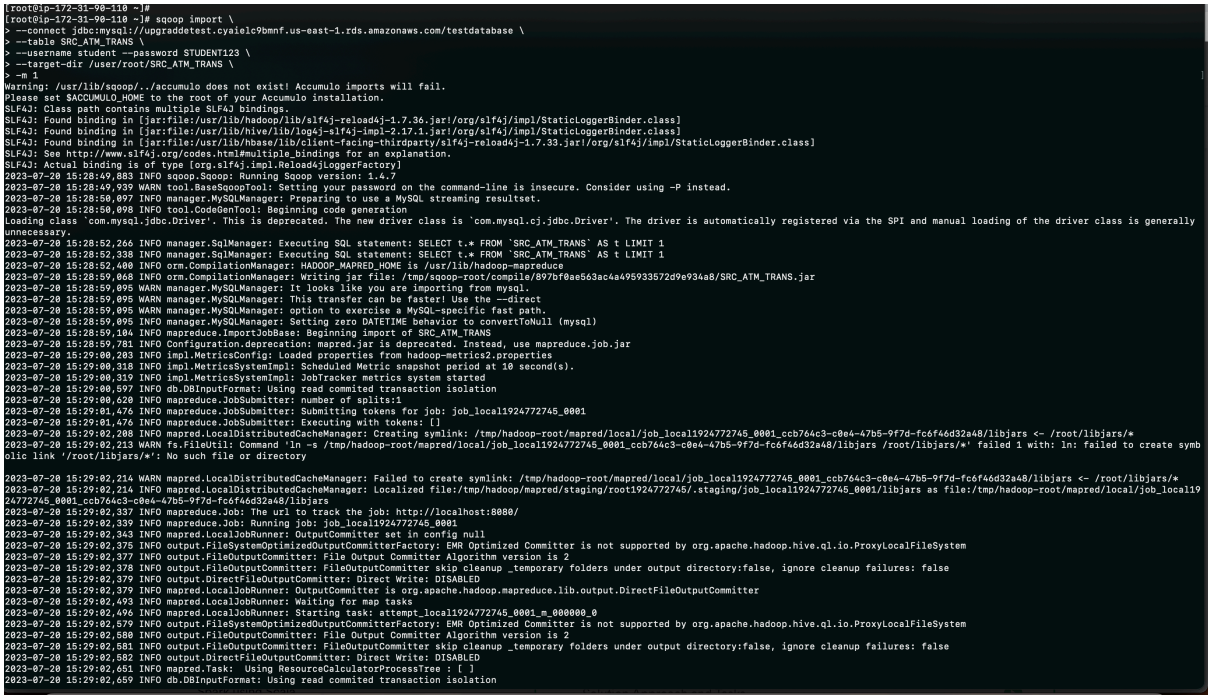


Step3: Installed JDBC connector using below commands.

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

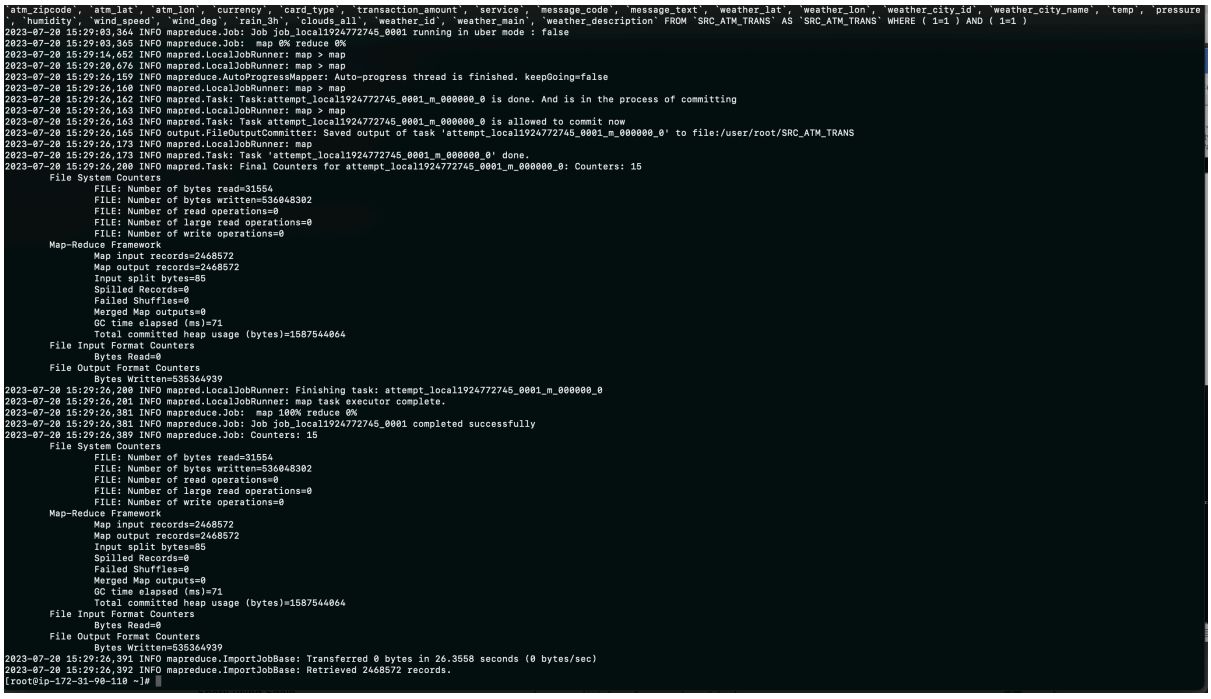
Step 3: Imported the data from the RDS using below sqoop command:

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



```
[root@ip-172-31-90-110 ~]#
[root@ip-172-31-90-110 ~]# sqoop import \
> --connect jdbc:mysql://upgraddetest.cyaielc9bmnf.us-east-1.rds.amazonaws.com/testdatabase \
> --table SRC_ATM_TRANS \
> --username student --password STUDENT123 \
> --target-dir /user/root/SRC_ATM_TRANS \
> -m 1
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hive/lib/log4j-slf4j-impl-2.17.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hbase/lib/client-facing-thirdparty/slf4j-reload4j-1.7.33.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Reload4jLoggerFactory]
2023-07-20 15:28:49,883 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
2023-07-20 15:28:49,939 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
2023-07-20 15:28:50,097 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
2023-07-20 15:28:50,098 INFO tool.CodeGenTool: Beginning code generation
Loading class `com.mysql.jdbc.Driver'. This is deprecated. The new driver class is `com.mysql.cj.jdbc.Driver'. The driver is automatically registered via the SPI and manual loading of the driver class is generally
unnecessary.
2023-07-20 15:28:52,266 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `SRC_ATM_TRANS` AS t LIMIT 1
2023-07-20 15:28:52,338 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `SRC_ATM_TRANS` AS t LIMIT 1
2023-07-20 15:28:52,400 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
2023-07-20 15:28:59,868 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-root/compile/897bf0ae563ac4a495933572d9e934a8/SRC_ATM_TRANS.jar
2023-07-20 15:28:59,095 WARN manager.MySQLManager: It looks like you are importing from mysql.
2023-07-20 15:28:59,095 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
2023-07-20 15:28:59,095 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
2023-07-20 15:28:59,095 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
2023-07-20 15:28:59,104 INFO mapreduce.ImportJobBase: Beginning import of SRC_ATM_TRANS
2023-07-20 15:28:59,781 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
2023-07-20 15:29:00,203 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2023-07-20 15:29:00,318 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2023-07-20 15:29:00,319 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2023-07-20 15:29:00,597 INFO db.DBInputFormat: Using read commited transaction isolation
2023-07-20 15:29:00,620 INFO mapreduce.JobSubmitter: number of splits:1
2023-07-20 15:29:01,476 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1924772745_0001
2023-07-20 15:29:01,476 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-07-20 15:29:02,208 INFO mapred.LocalDistributedCacheManager: Creating symlink: /tmp/hadoop-root/mapred/local/job_local1924772745_0001_ccb764c3-c0e4-47b5-9f7d-fc6f46d32a48/libjars <- /root/libjars/*
2023-07-20 15:29:02,213 WARN fs.FileUtil: Command 'ln -s /tmp/hadoop-root/mapred/local/job_local1924772745_0001_ccb764c3-c0e4-47b5-9f7d-fc6f46d32a48/libjars /root/libjars/*' failed 1 with: ln: failed to create symb
olic link '/root/libjars/*': No such file or directory

2023-07-20 15:29:02,214 WARN mapred.LocalDistributedCacheManager: Failed to create symlink: /tmp/hadoop-root/mapred/local/job_local1924772745_0001_ccb764c3-c0e4-47b5-9f7d-fc6f46d32a48/libjars <- /root/libjars/*
2023-07-20 15:29:02,214 INFO mapred.LocalDistributedCacheManager: Localized file:/tmp/hadoop/mapred/staging/root1924772745/.staging/job_local1924772745_0001/libjars as file:/tmp/hadoop-root/mapred/local/job_local19
24772745_0001_ccb764c3-c0e4-47b5-9f7d-fc6f46d32a48/libjars
2023-07-20 15:29:02,337 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2023-07-20 15:29:02,339 INFO mapreduce.Job: Running job: job_local1924772745_0001
2023-07-20 15:29:02,343 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2023-07-20 15:29:02,375 INFO output.FileSystemOptimizedOutputCommitterFactory: EMR Optimized Committer is not supported by org.apache.hadoop.hive.ql.io.ProxyLocalFileSystem
2023-07-20 15:29:02,377 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2023-07-20 15:29:02,378 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2023-07-20 15:29:02,379 INFO output.DirectFileOutputCommitter: Direct Write: DISABLED
2023-07-20 15:29:02,379 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.DirectFileOutputCommitter
2023-07-20 15:29:02,493 INFO mapred.LocalJobRunner: Waiting for map tasks
2023-07-20 15:29:02,496 INFO mapred.LocalJobRunner: Starting task: attempt_local1924772745_0001_m_000000_0
2023-07-20 15:29:02,579 INFO output.FileSystemOptimizedOutputCommitterFactory: EMR Optimized Committer is not supported by org.apache.hadoop.hive.ql.io.ProxyLocalFileSystem
2023-07-20 15:29:02,580 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2023-07-20 15:29:02,581 INFO output.FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2023-07-20 15:29:02,582 INFO output.DirectFileOutputCommitter: Direct Write: DISABLED
2023-07-20 15:29:02,651 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2023-07-20 15:29:02,659 INFO db.DBInputFormat: Using read commited transaction isolation
```



```
`atm_zipcode`, `atm_lat`, `atm_lon`, `currency`, `card_type`, `transaction_amount`, `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` FROM `SRC_ATM_TRANS` AS `SRC_ATM_TRANS` WHERE ( 1=1 ) AND ( 1=1 )
2023-07-20 15:29:03,364 INFO mapreduce.Job: Job job_local1924772745_0001 running in uber mode : false
2023-07-20 15:29:03,365 INFO mapreduce.Job:  map 0% reduce 0%
2023-07-20 15:29:14,652 INFO mapred.LocalJobRunner: map > map
2023-07-20 15:29:20,676 INFO mapred.LocalJobRunner: map > map
2023-07-20 15:29:26,159 INFO mapreduce.AutoProgressMapper: Auto-progress thread is finished. keepGoing=false
2023-07-20 15:29:26,160 INFO mapred.LocalJobRunner: map > map
2023-07-20 15:29:26,162 INFO mapred.Task: Task:attempt_local1924772745_0001_m_000000_0 is done. And is in the process of committing
2023-07-20 15:29:26,163 INFO mapred.LocalJobRunner: map > map
2023-07-20 15:29:26,163 INFO mapred.Task: Task attempt_local1924772745_0001_m_000000_0 is allowed to commit now
2023-07-20 15:29:26,165 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1924772745_0001_m_000000_0' to file:/user/root/SRC_ATM_TRANS
2023-07-20 15:29:26,173 INFO mapred.LocalJobRunner: map
2023-07-20 15:29:26,173 INFO mapred.Task: Task 'attempt_local1924772745_0001_m_000000_0' done.
2023-07-20 15:29:26,200 INFO mapred.Task: Final Counters for attempt_local1924772745_0001_m_000000_0: Counters: 15
        File System Counters
                FILE: Number of bytes read=31554
                FILE: Number of bytes written=536048302
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
        Map-Reduce Framework
                Map input records=2468572
                Map output records=2468572
                Input split bytes=85
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=71
                Total committed heap usage (bytes)=1587544064
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=535364939
2023-07-20 15:29:26,200 INFO mapred.LocalJobRunner: Finishing task: attempt_local1924772745_0001_m_000000_0
2023-07-20 15:29:26,201 INFO mapred.LocalJobRunner: map task executor complete.
2023-07-20 15:29:26,381 INFO mapreduce.Job:  map 100% reduce 0%
2023-07-20 15:29:26,381 INFO mapreduce.Job: Job job_local1924772745_0001 completed successfully
2023-07-20 15:29:26,389 INFO mapreduce.Job: Counters: 15
        File System Counters
                FILE: Number of bytes read=31554
                FILE: Number of bytes written=536048302
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
        Map-Reduce Framework
                Map input records=2468572
                Map output records=2468572
                Input split bytes=85
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=71
                Total committed heap usage (bytes)=1587544064
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=535364939
2023-07-20 15:29:26,391 INFO mapreduce.ImportJobBase: Transferred 0 bytes in 26.3558 seconds (0 bytes/sec)
2023-07-20 15:29:26,392 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
[root@ip-172-31-90-110 ~]#
```

Step 4: Checked the data from the location using below command to ensure it is imported.

hadoop fs -ls /user/root/SRC_ATM_TRANS