

**Task 2.** Use Sqoop command to ingest the data from RDS into the HBase Table.

**Solution:** Here's the step by step procedure for achieving the above task:

**Step1:** We will first download mysql connector and unzip the file.

`tar -xzvf mysql-connector-j-9.1.0.tar.gz`

```
[hadoop@ip-172-31-32-108 ~]$ sudo wget https://cdn.mysql.com//Downloads/Connector-J/mysql-connector-j-9.1.0.tar.gz
--2025-01-07 05:28:51-- https://cdn.mysql.com//Downloads/Connector-J/mysql-connector-j-9.1.0.tar.gz
Resolving cdn.mysql.com (cdn.mysql.com)... 184.25.42.21, 2600:1408:c400:188d::1d68, 2600:1408:c400:1884::1d68
Connecting to cdn.mysql.com (cdn.mysql.com)|184.25.42.21|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 4485702 (4.3M) [application/x-tar-gz]
Saving to: 'mysql-connector-j-9.1.0.tar.gz'

mysql-connector-j-9.1.0.tar.gz      100%[=====] 4.28M --.-KB/s  in 0.1s

2025-01-07 05:28:51 (34.3 MB/s) - 'mysql-connector-j-9.1.0.tar.gz' saved [4485702/4485702]

[hadoop@ip-172-31-32-108 ~]$ tar -xzvf mysql-connector-j-9.1.0.tar.gz
mysql-connector-j-9.1.0/
mysql-connector-j-9.1.0/src/
mysql-connector-j-9.1.0/src/build/
mysql-connector-j-9.1.0/src/build/java/
mysql-connector-j-9.1.0/src/build/java/documentation/
mysql-connector-j-9.1.0/src/build/java/instrumentation/
mysql-connector-j-9.1.0/src/build/misc/
mysql-connector-j-9.1.0/src/build/misc/debian.in/
mysql-connector-j-9.1.0/src/build/misc/debian.in/source/
mysql-connector-j-9.1.0/src/demo/
mysql-connector-j-9.1.0/src/demo/java/
mysql-connector-j-9.1.0/src/demo/java/demo/
mysql-connector-j-9.1.0/src/demo/java/demo/x/
mysql-connector-j-9.1.0/src/demo/java/demo/x/devapi/
mysql-connector-j-9.1.0/src/generated/
mysql-connector-j-9.1.0/src/generated/java/
mysql-connector-j-9.1.0/src/generated/java/com/
mysql-connector-j-9.1.0/src/generated/java/com/mysql/
mysql-connector-j-9.1.0/src/generated/java/com/mysql/cj/
```

**Step 2:** `sqoop list-databases --connect jdbc:mysql://database-1.crkohtdqsjvo.us-east-1.rds.amazonaws.com --username admin --password *****`

```
[hadoop@ip-172-31-32-108 ~]$ sqoop list-databases --connect jdbc:mysql://database-1.crkohtdqsjvo.us-east-1.rds.amazonaws.com --username admin --password *****
Warning: /usr/lib/sqoop/./accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hive/lib/log4j-slf4j-impl-2.17.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hbase/lib/client-facing-thirdparty/log4j-slf4j-impl-2.17.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Reload4jLoggerFactory]
2025-01-07 05:53:09,366 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
2025-01-07 05:53:09,404 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
2025-01-07 05:53:09,534 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
Loading class 'com.mysql.jdbc.Driver'. This is deprecated. The new driver class is 'com.mysql.cj.jdbc.Driver'. The driver is automatically registered via the SPI and manual loading of the driver class is generally unnecessary.
mysql
information_schema
performance_schema
sys
assignment
[hadoop@ip-172-31-32-108 ~]$
```

**Step 3:** `sqoop import --connect jdbc:mysql://database-1.crkohtdqsjvo.us-east-1.rds.amazonaws.com/assignment --username admin --password ***** --table TaxiTripData --target-dir /TaxiTripDataHbase --m 1`

```
Try --help for usage instructions.
[hadoop@ip-172-31-32-108 ~]$ sqoop import --connect jdbc:mysql://database-1.crkohtdqsjvo.us-east-1.rds.amazonaws.com/assignment --username admin --password ***** --table TaxiTripData --target-dir /TaxiTripDataHbase --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/log4j-slf4j-impl-2.17.2.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]
2025-01-07 06:02:05,986 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
2025-01-07 06:02:06,028 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
2025-01-07 06:02:06,189 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
2025-01-07 06:02:06,189 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.
2025-01-07 06:02:07,118 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM 'TaxiTripData' AS t LIMIT 1
2025-01-07 06:02:07,197 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM 'TaxiTripData' AS t LIMIT 1
2025-01-07 06:02:07,218 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
2025-01-07 06:02:13,861 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hadoop/compile/d5899caf4a53970835b47c2ee400f955/TaxiTripData.jar
2025-01-07 06:02:13,880 WARN manager.MySQLManager: It looks like you are importing from mysql.
2025-01-07 06:02:13,880 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
2025-01-07 06:02:13,880 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
2025-01-07 06:02:13,881 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
```

Step 4: Now we can see that sqoop has ingested the data from RDS into Hbase table.

```
2025-01-07 06:02:19,854 INFO impl.YarnClientImpl: Submitted application application_1736226425433_0001
2025-01-07 06:02:19,905 INFO mapreduce.Job: The url to track the job: http://ip-172-31-32-108.ec2.internal:20888/proxy/application_1736226425433_0001/
2025-01-07 06:02:19,906 INFO mapreduce.Job: Running job: job_1736226425433_0001
2025-01-07 06:02:29,066 INFO mapreduce.Job: Job job_1736226425433_0001 running in uber mode : false
2025-01-07 06:02:29,067 INFO mapreduce.Job: map 0% reduce 0%
2025-01-07 06:05:42,070 INFO mapreduce.Job: map 100% reduce 0%
2025-01-07 06:05:43,079 INFO mapreduce.Job: Job job_1736226425433_0001 completed successfully
2025-01-07 06:05:43,159 INFO mapreduce.Job: Counters: 33
  File System Counters
    FILE: Number of bytes read=0
    FILE: Number of bytes written=332156
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=85
    HDFS: Number of bytes written=1966321996
    HDFS: Number of read operations=6
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
    HDFS: Number of bytes read erasure-coded=0
  Job Counters
    Launched map tasks=1
    Other local map tasks=1
    Total time spent by all maps in occupied slots (ms)=293227008
    Total time spent by all reduces in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=190903
    Total vcore-milliseconds taken by all map tasks=190903
    Total megabyte-milliseconds taken by all map tasks=293227008
  Map-Reduce Framework
    Map input records=18880595
    Map output records=18880595
    Input split bytes=85
    Spilled Records=0
    Failed Shuffles=0
    Merged Map outputs=0
    GC time elapsed (ms)=1139
    CPU time spent (ms)=204400
    Physical memory (bytes) snapshot=469876736
    Virtual memory (bytes) snapshot=3196207104
    Total committed heap usage (bytes)=306184192
    Peak Map Physical memory (bytes)=469876736
    Peak Map Virtual memory (bytes)=3196207104
  File Input Format Counters
    Bytes Read=0
  File Output Format Counters
    Bytes Written=1966321996
2025-01-07 06:05:43,164 INFO mapreduce.ImportJobBase: Transferred 1.8313 GB in 208.5483 seconds (8.9918 MB/sec)
2025-01-07 06:05:43,166 INFO mapreduce.ImportJobBase: Retrieved 18880595 records.
[hadoop@ip-172-31-32-108 ~]$
```