

HBase from EMR

Creating table in HBase

```
[root@ip-172-31-11-118:~]# sudo -i
[hadoop@ip-172-31-11-118 ~]$ sudo -i
EEEEEEEEEEEEEEEEEEEE MMMMMMMM MMMMMMMM RRRRRRRRRRRRRR
E:::EEEEEEEEEEEEEEEE M:::MM M:::MM R:::RRRRRRRRRRRR
EE:::EEEEEEEEEEEEEEEE M:::MM M:::MM R:::RRRRRRRRRRRR
E:::E EEEEE M:::MM M:::MM M:::MM RR:::R R:::R
E:::E EEEEE M:::MM M:::MM M:::MM RR:::R R:::R
E:::EEEEEEEEEEEE M:::MM M:::MM M:::MM R:::RRRRRRRRRR
E:::EEEEEEEEEEEE M:::MM M:::MM M:::MM R:::RRRRRRRRRR
E:::EEEEEEEEEEEE M:::MM M:::MM M:::MM R:::RRRRRRRRRR
E:::E M:::MM M:::MM M:::MM R:::R R:::R
E:::E EEEEE M:::MM M:::MM M:::MM R:::R R:::R
EE:::EEEEEEEEEEEE M:::MM M:::MM M:::MM R:::R R:::R
E:::EEEEEEEEEEEE M:::MM M:::MM RR:::R R:::R
EEEEEEEEEEEEEEEEEEEE MMMMMMMM MMMMMMMM RRRRRRR RRRRRR

[root@ip-172-31-11-118 ~]# hbase shell
HBase Shell
Use "help" to get list of supported commands.
Use "exit" to quit this interactive shell.
Version 1.4.13, rUnknown, Fri Apr 17 15:18:24 UTC 2020

hbase(main):001:0> create 'yellow_tripdata', 'data'
0 row(s) in 2.5940 seconds

=> Hbase::Table - yellow_tripdata
hbase(main):002:0> Describe 'yellow_tripdata'
NoMethodError: undefined method `Describe' for #<Object:0x5002fde9>

hbase(main):003:0> describe 'yellow_tripdata'
Table yellow_tripdata is ENABLED
yellow_tripdata
COLUMN FAMILIES DESCRIPTION
{NAME => 'data', BLOOMFILTER => 'ROW', VERSIONS => '1', IN_MEMORY => 'false', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', TTL => 'FOREVER', COMPRESSION => 'NONE', MIN_VERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}
1 row(s) in 0.0320 seconds

hbase(main):004:0>
```

```

hadoop@ip-172-31-11-210:~
at org.jruby.Main.run(Main.java:224)
at org.jruby.Main.run(Main.java:208)
at org.jruby.Main.main(Main.java:188)
log4j:ERROR Either File or DatePattern options are not set for appender [DRFAS].
HBase Shell
Use "help" to get list of supported commands.
Use "exit" to quit this interactive shell.
Version 1.4.13, rUnknown, Fri Apr 17 15:18:24 UTC 2020

hbase(main):001:0> create 'yellow_tripdata', 'data'
0 row(s) in 2.5490 seconds

=> Hbase::Table - yellow_tripdata
hbase(main):002:0> describe 'yellow_tripdata'
Table yellow_tripdata is ENABLED
yellow_tripdata
COLUMN FAMILIES DESCRIPTION
{NAME => 'data', BLOOMFILTER => 'ROW', VERSIONS => '1', IN_MEMORY => 'false', KE
EP DELETED CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', TTL => 'FOREVER', CO
MPRESSION => 'NONE', MIN_VERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '65
536', REPLICATION_SCOPE => '0'}
1 row(s) in 0.0350 seconds

hbase(main):003:0>

```

wget <https://repo1.maven.org/maven2/com/mysql/mysql-connector-j/8.0.33/mysql-connector-j-8.0.33.jar>

```
hadoop@ip-172-31-11-210:~$ wget https://dev.mysql.com/get/Downloads/Connector-J/mysql-connector-java-8.0.33.tar.gz
--2025-02-02 07:26:29-- https://dev.mysql.com/get/Downloads/Connector-J/mysql-connector-java-8.0.33.tar.gz
Resolving dev.mysql.com (dev.mysql.com)... 23.49.176.249, 2600:1408:c400:1788::2e31, 2600:1408:c400:178d::2e31
Connecting to dev.mysql.com (dev.mysql.com)|23.49.176.249|:443... connected.
HTTP request sent, awaiting response... 404 Not Found
2025-02-02 07:26:30 ERROR 404: Not Found.

[hadoop@ip-172-31-11-210 ~]$ wget https://repo1.maven.org/maven2/com/mysql/mysql-connector-j/8.0.33/mysql-connector-j-8.0.33.jar
--2025-02-02 07:27:02-- https://repo1.maven.org/maven2/com/mysql/mysql-connector-j/8.0.33/mysql-connector-j-8.0.33.jar
Resolving repo1.maven.org (repo1.maven.org)... 146.75.32.209, 2a04:4e42:78::209
Connecting to repo1.maven.org (repo1.maven.org)|146.75.32.209|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 2481560 (2.4M) [application/java-archive]
Saving to: 'mysql-connector-j-8.0.33.jar'

100%[=====] 2,481,560 --.-K/s in 0.01s

2025-02-02 07:27:02 (172 MB/s) - 'mysql-connector-j-8.0.33.jar' saved [2481560/2481560]

[hadoop@ip-172-31-11-210 ~]$
```

After downloading, move the JAR file to Sqoop's library directory to ensure Sqoop can access the MySQL driver:

`sudo mv mysql-connector-j-8.0.33.jar /usr/lib/sqoop/lib/`

```
hadoop@ip-172-31-11-210:~$ sudo mv mysql-connector-j-8.0.33.jar /usr/lib/sqoop/lib/
Connecting to dev.mysql.com (dev.mysql.com)|23.49.176.249|:443... connected.
HTTP request sent, awaiting response... 404 Not Found
2025-02-02 07:26:30 ERROR 404: Not Found.

[hadoop@ip-172-31-11-210 ~]$ wget https://repo1.maven.org/maven2/com/mysql/mysql-connector-j/8.0.33/mysql-connector-j-8.0.33.jar
--2025-02-02 07:27:02-- https://repo1.maven.org/maven2/com/mysql/mysql-connector-j/8.0.33/mysql-connector-j-8.0.33.jar
Resolving repo1.maven.org (repo1.maven.org)... 146.75.32.209, 2a04:4e42:78::209
Connecting to repo1.maven.org (repo1.maven.org)|146.75.32.209|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 2481560 (2.4M) [application/java-archive]
Saving to: 'mysql-connector-j-8.0.33.jar'

100%[=====] 2,481,560 --.-K/s in 0.01s

2025-02-02 07:27:02 (172 MB/s) - 'mysql-connector-j-8.0.33.jar' saved [2481560/2481560]

[hadoop@ip-172-31-11-210 ~]$ ls /usr/lib/sqoop/lib/ | grep mysql
mysql-connector-j-8.0.33.jar
[hadoop@ip-172-31-11-210 ~]$
```

Sqoop import

Since table lacks a primary key, using a **numerical column** as a split key: PULocationID (Pickup Location ID).

`sqoop import \`

```
--connect jdbc:mysql://mysqladb.cyhl71hae9td.us-east-1.rds.amazonaws.com/mysqladb \

--username admin \

--password admin123 \

--table yellow_tripdata \

--hbase-table yellow_tripdata \

--column-family data \

--hbase-create-table \

--split-by PULocationID
```

```
fast path.
25/02/02 07:29:56 INFO manager.MySQLManager: Setting zero DATETIME behavior to c
onvertToNull (mysql)
25/02/02 07:29:56 ERROR tool.ImportTool: Import failed: No primary key could be
found for table yellow_tripdata. Please specify one with --split-by or perform a
sequential import with '-m 1'.
[hadoop@ip-172-31-11-210 ~]$ sqoop import \
> --connect jdbc:mysql://mysqladb.cyhl71hae9td.us-east-1.rds.amazonaws.com/mysqladb \
> --username admin \
> --password admin123 \
> --table yellow_tripdata \
> --hbase-table yellow_tripdata \
> --column-family data \
> --hbase-create-table \
> --split-by PULocationID
Warning: /usr/lib/sqoop/./accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
25/02/02 07:34:10 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.10.jar!/
org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/aws/redshift/jdbc/redshift-jdbc42-1
.2.37.1061.jar!/org/slf4j/impl/StaticLoggerBinder.class]

Total time spent by all reduces in occupied slots (ms)=0
Total time spent by all map tasks (ms)=2749831
Total vcore-milliseconds taken by all map tasks=2749831
Total megabyte-milliseconds taken by all map tasks=4223740416
Map-Reduce Framework
  Map input records=18880595
  Map output records=18880595
  Input split bytes=485
  Spilled Records=0
  Failed Shuffles=0
  Merged Map outputs=0
  GC time elapsed (ms)=21738
  CPU time spent (ms)=1187740
  Physical memory (bytes) snapshot=2926821376
  Virtual memory (bytes) snapshot=13238489088
  Total committed heap usage (bytes)=2439512064
File Input Format Counters
  Bytes Read=0
File Output Format Counters
  Bytes Written=0
25/02/02 08:00:59 INFO mapreduce.ImportJobBase: Transferred 0 bytes in 1,602.730
6 seconds (0 bytes/sec)
25/02/02 08:00:59 INFO mapreduce.ImportJobBase: Retrieved 18880595 records.
[hadoop@ip-172-31-11-210 ~]$
```

```
scan 'yellow_tripdata', {LIMIT => 2}
```

```
hadoop@ip-172-31-11-210~  
109          column=data:trip_distance, timestamp=1738482274328, value=  
0.0  
10 row(s) in 0.4440 seconds  
  
hbase(main):002:0> scan 'yellow_tripdata', {LIMIT => 2}  
ROW          COLUMN+CELL  
1            column=data:DOLocationID, timestamp=1738481960838, value=1  
1            column=data:RatecodeID, timestamp=1738481960838, value=5  
1            column=data:VendorID, timestamp=1738481960838, value=2  
1            column=data:extra, timestamp=1738481960838, value=0.0  
1            column=data:fare_amount, timestamp=1738481960838, value=65  
.0  
1            column=data:improvement_surcharge, timestamp=1738481960838  
, value=0.3  
1            column=data:mta_tax, timestamp=1738481960838, value=0.5  
1            column=data:passenger_count, timestamp=1738481960838, valu  
e=2  
1            column=data:payment_type, timestamp=1738481960838, value=1  
1            column=data:store_and_fwd_flag, timestamp=1738481960838, v  
alue=N  
1            column=data:tip_amount, timestamp=1738481960838, value=0.0  
1            column=data:tolls_amount, timestamp=1738481960838, value=0  
.0  
1            column=data:total_amount, timestamp=1738481960838, value=6
```