# Ingestion Task

In this task of MapReduce - Programming Assignment, we have created a HBase table then written a Sqoop command to ingest data from RDS into the HBase table. Below are the steps and code we have used to achieve this task.

**<u>Step 1</u>: Open HBase Shell in EMR:**

**Code:**

*Hbase shell*

**Explanation:**
HBase has a shell that we may use to interact with it, "hbase shell" can be used to launch the interactive HBase shell.

```
[[hadoop@ip-172-31-14-109 ~]$ hbase shell
```

**<u>Step 2</u>: Create a table in hbase:**

**Code:**

*create 'trip_record', 'trip_details'*

**Explanation:**
Using the create command, we can construct a table; in this case, we must enter the table name and the name of the Column Family. The HBase shell command for creating a table is displayed below.

create '<table name>','<column family>'

Here in our case, our table name is *'trip_record'* and the column family name is *'trip_details'*

```
hbase(main):017:0> create 'trip_record', 'trip_details'
0 row(s) in 1.2610 seconds

=> Hbase::Table - trip_record
```

Check whether table is create or not

**Code**

*list*

**Explanation:**
The command *list* is used to display a list of each and every table in HBase.

```
hbase(main):018:0> list
TABLE
trip_record
1 row(s) in 0.0070 seconds

=> ["trip_record"]
```

**<u>Step 3</u>: Sqoop import command:**

**Code**

*sqoop import --connect jdbc:mysql://mapreduceassignment.cjipwhexa9kv.us-east-1.rds.amazonaws.com:3306/MapReduceAssignment --driver com.mysql.jdbc.Driver --username admin --password nikhil02022 --table trip_record --hbase-create-table --hbase-table trip_record --column-family trip_details --hbase-row-key VendorID,tpep_pickup_datetime,tpep_dropoff_datetime --split-by VendorID*

**Explanation:**

The AWS RDS table data is imported into Hbase using the Sqoop tool "import"
Here,

- **--connect <jdbc-uri>** : Indicate the JDBC connect string.
- **--driver <class-name>** : To choose the appropriate JDBC driver class manually
- **--username <username>** : Choose a username for authentication
- **--password <password>** : set a password for authentication
- **--table <table-name>** : Read the table
- **--hbase-create-table** : Create any missing HBase tables if provided.
- **--hbase-table <table-name>** : Sets an HBase table as the target rather than HDFS
- **--column-family <family>** : Establishes the import's target column family
- **--hbase-row-key <col>** : Which input column should be used as the row key.
- **--split-by <column-name>** : The table's column used to divide up work units. Use with the —autoreset-to-one-mapper option is not permitted.

```
[hadoop@ip-172-31-2-78 mysql-connector-java-8.0.25]$ sqoop import --connect jdbc:mysql://mapreduceassignment.cjipwhexa9kv.us-east-1.rds.amazonaws.com:3306/MapReduceAssignment --driver com.mysql.jdbc.Driver --userna
me admin --password nikhil02022 --table trip_record  --hbase-create-table  --hbase-table trip_record --column-family trip_details --hbase-row-key VendorID,tpep_pickup_datetime,tpep_dropoff_datetime --split-by Vendo
rID
```

Sqoop Command Status:

```
[hadoop@ip-172-31-2-78 mysql-connector-java-8.0.25]$ sqoop import --connect jdbc:mysql://mapreduceassignment.cjipwhexa9kv.us-east-1.rds.amazonaws.com:3306/MapReduceAssignment --driver com.mysql.jdbc.Driver --username admin --password nikhil02022 --table trip_record  --hbase-create-table  --hbase-table trip_record --column-family trip_details --hbase-row-key VendorID,tpep_pickup_datetime,tpep_dropoff_datetime --split-by VendorID
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-log4j12-1.7.25.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: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
22/12/21 07:41:38 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
22/12/21 07:41:38 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
22/12/21 07:41:38 WARN sqoop.ConnFactory: Parameter --driver is set to an explicit driver however appropriate connection manager is not being set (via --connection-manager). Sqoop is going to fall back to org.apache.sqoop.manager.GenericJdbcManager. Please specify explicitly which connection manager should be used next time.
22/12/21 07:41:38 INFO manager.SqlManager: Using default fetchSize of 1000
22/12/21 07:41:38 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.
22/12/21 07:41:39 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM trip_record AS t WHERE 1=0
22/12/21 07:41:39 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM trip_record AS t WHERE 1=0
22/12/21 07:41:39 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-hadoop/compile/49e5900f09d31dcb009650573014b171/trip_record.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
22/12/21 07:41:45 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hadoop/compile/49e5900f09d31dcb009650573014b171/trip_record.jar
22/12/21 07:41:45 INFO mapreduce.ImportJobBase: Beginning import of trip_record
22/12/21 07:41:46 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
22/12/21 07:41:46 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM trip_record AS t WHERE 1=0
22/12/21 07:41:46 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
22/12/21 07:41:49 WARN mapreduce.TableMapReduceUtil: The addDependencyJars(Configuration, Class<?>...) method has been deprecated since it is easy to use incorrectly. Most users should rely on addDependencyJars(Job) instead. See HBASE-8386 for more details.
22/12/21 07:41:49 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-2-78.ec2.internal/172.31.2.78:8032
22/12/21 07:41:49 INFO client.AHSProxy: Connecting to Application History server at ip-172-31-2-78.ec2.internal/172.31.2.78:10200
22/12/21 07:42:02 INFO db.DBInputFormat: Using read commited transaction isolation
22/12/21 07:42:02 INFO db.DataDrivenDBInputFormat: BoundingValsQuery: SELECT MIN(VendorID), MAX(VendorID) FROM trip_record
22/12/21 07:42:36 INFO db.IntegerSplitter: Split size: 0; Num splits: 4 from: 1 to: 2
22/12/21 07:42:36 INFO mapreduce.JobSubmitter: number of splits:2
```

MR Job Status at 50% map task completed:

```
ob:  map 50% reduce 0%
```

MR Job Status at 100% map task completed:

```
INFO mapreduce.Job:  map 100% reduce 0%
```

Sqoop import successful completion:

```
INFO mapreduce.Job: Job job_1671389176913_0003 completed successfully
INFO mapreduce.Job: Counters: 30
m Counters
LE: Number of bytes read=0
LE: Number of bytes written=521324
LE: Number of read operations=0
LE: Number of large read operations=0
LE: Number of write operations=0
DFS: Number of bytes read=213
DFS: Number of bytes written=0
DFS: Number of read operations=2
DFS: Number of large read operations=0
DFS: Number of write operations=0
ers
unched map tasks=3
ther local map tasks=3
otal time spent by all maps in occupied slots (ms)=156474048
otal time spent by all reduces in occupied slots (ms)=0
otal time spent by all map tasks (ms)=3259876
otal vcore-milliseconds taken by all map tasks=3259876
otal megabyte-milliseconds taken by all map tasks=5007169536
 Framework
ap input records=18880595
ap output records=18880595
nput split bytes=213
pilled Records=0
ailed Shuffles=0
erged Map outputs=0
 time elapsed (ms)=25774
PU time spent (ms)=1292930
hysical memory (bytes) snapshot=1616650240
irtual memory (bytes) snapshot=6715441152
otal committed heap usage (bytes)=1285029888
 Format Counters
ytes Read=0
t Format Counters
ytes Written=0
INFO mapreduce.ImportJobBase: Transferred 0 bytes in 1,849.3161 seconds (0 bytes/sec)
INFO mapreduce.ImportJobBase: Retrieved 18880595 records.
```

**<u>Step 4</u>: Verify data in HBase Table:**

Printed first row data in hbase

**Code**

*scan 'trip_record'*

**Explanation:**

The **scan** command is used to view the data in HTable.

```
hbase(main):001:0> scan 'trip_record'
ROW                                                  COLUMN+CELL
 1_2017-01-01 00:00:02.0_2017-01-01 00:03:50.0       column=trip_details:DOLocationID, timestamp=1671608899845, value=48
 1_2017-01-01 00:00:02.0_2017-01-01 00:03:50.0       column=trip_details:PULocationID, timestamp=1671608899845, value=48
 1_2017-01-01 00:00:02.0_2017-01-01 00:03:50.0       column=trip_details:RatecodeID, timestamp=1671608899845, value=1
 1_2017-01-01 00:00:02.0_2017-01-01 00:03:50.0       column=trip_details:airport_fee, timestamp=1671608899845, value=0.0
 1_2017-01-01 00:00:02.0_2017-01-01 00:03:50.0       column=trip_details:congestion_surcharge, timestamp=1671608899845, value=0.0
 1_2017-01-01 00:00:02.0_2017-01-01 00:03:50.0       column=trip_details:extra, timestamp=1671608899845, value=0.5
 1_2017-01-01 00:00:02.0_2017-01-01 00:03:50.0       column=trip_details:fare_amount, timestamp=1671608899845, value=4.0
 1_2017-01-01 00:00:02.0_2017-01-01 00:03:50.0       column=trip_details:improvement_surcharge, timestamp=1671608899845, value=0.3
 1_2017-01-01 00:00:02.0_2017-01-01 00:03:50.0       column=trip_details:mta_tax, timestamp=1671608899845, value=0.5
 1_2017-01-01 00:00:02.0_2017-01-01 00:03:50.0       column=trip_details:passenger_count, timestamp=1671608899845, value=1
 1_2017-01-01 00:00:02.0_2017-01-01 00:03:50.0       column=trip_details:payment_type, timestamp=1671608899845, value=2
 1_2017-01-01 00:00:02.0_2017-01-01 00:03:50.0       column=trip_details:store_and_fwd_flag, timestamp=1671608899845, value=N
 1_2017-01-01 00:00:02.0_2017-01-01 00:03:50.0       column=trip_details:tip_amount, timestamp=1671608899845, value=0.00
 1_2017-01-01 00:00:02.0_2017-01-01 00:03:50.0       column=trip_details:tolls_amount, timestamp=1671608899845, value=0.00
 1_2017-01-01 00:00:02.0_2017-01-01 00:03:50.0       column=trip_details:total_amount, timestamp=1671608899845, value=5.30
 1_2017-01-01 00:00:02.0_2017-01-01 00:03:50.0       column=trip_details:trip_distance, timestamp=1671608899845, value=0.50
 1_2017-01-01 00:00:03.0_2017-01-01 00:06:58.0       column=trip_details:DOLocationID, timestamp=1671608898944, value=161
 1_2017-01-01 00:00:03.0_2017-01-01 00:06:58.0       column=trip_details:PULocationID, timestamp=1671608898944, value=162
 1_2017-01-01 00:00:03.0_2017-01-01 00:06:58.0       column=trip_details:RatecodeID, timestamp=1671608898944, value=1
 1_2017-01-01 00:00:03.0_2017-01-01 00:06:58.0       column=trip_details:airport_fee, timestamp=1671608898944, value=0.0
 1_2017-01-01 00:00:03.0_2017-01-01 00:06:58.0       column=trip_details:congestion_surcharge, timestamp=1671608898944, value=0.0
```

**Submitted By:**

Member 1: Nikhil Dhiman (nikhildhiman3644@gmail.com)
Member 2: Rohan kulkarni (rohan.kulkarni951@gmail.com)
Member 3: Mahesh Kumar Patra (maheshkumarpatra5@gmail.com)