Mid-Submission – LOGIC

Explanation of the solution to the batch layer problem

Mentioned below is a summary of batch layer problem solution:

Step 1: Basic setup for Project as mentioned in project brief:

a. EMR cluster setup with required apps and configuration.
b. Access EMR cluster using AWS console/ssh.
c. Setup the project directory.
d. Download card_transactions.csv file to cluster's home directory.
e. Upload the above csv file to HDFS.
f. Data ingestion from RDS to HDFS.

Step 2: Address task 1-4 for Mid submission of capstone project:

Individual task approaches are as mentioned below:

Task 1: Load transaction history data in a NoSQL database

1. Go to 'hive' shell and go to project database(ccfd_hive_db);
2. Session configuration – set hive database parameters are appropriate.
3. Table creation – create table - external table named 'card_transactions_ext' and ORC format table named 'card_transactions_orc'
4. Data load – load data into 'card_transactions_orc' and convert timestamp in appropriate format.
5. Verify the data year – verify the year from transaction date column.
6. Hive-HBase integrated table – create a table named 'card_transactions_hbase' in hive that is backed by HBase storage
7. Data load into hive-HBase table – Load data and verify the count of rows.

Task 2: Data Ingestion from RDS to Hadoop

1. Sqoop import job – Import data from RDS to Hadoop using sqoop for 'member_score' and 'card_member' tables.
2. Table creation – Create external and ORC (to optimize space for better performance) format table for 'card_member' and 'member_score'.

3. Data Loading – use external table for loading data into ORC tables and validate rows' count.

Task 3: Create hive-HBase integrated table for lookup.

1. Fire a create table query in hive that is backed by HBase storage.
2. Verify the table details in HBase.

Task 4: Load data in lookup table.

1. Create and load 'ranked_card_transactions_orc' and 'card_ucl_orc' tables.
2. Load and verify lookup data.

Given below are the screenshot highlighting the code and commands executed to achieve the above tasks:

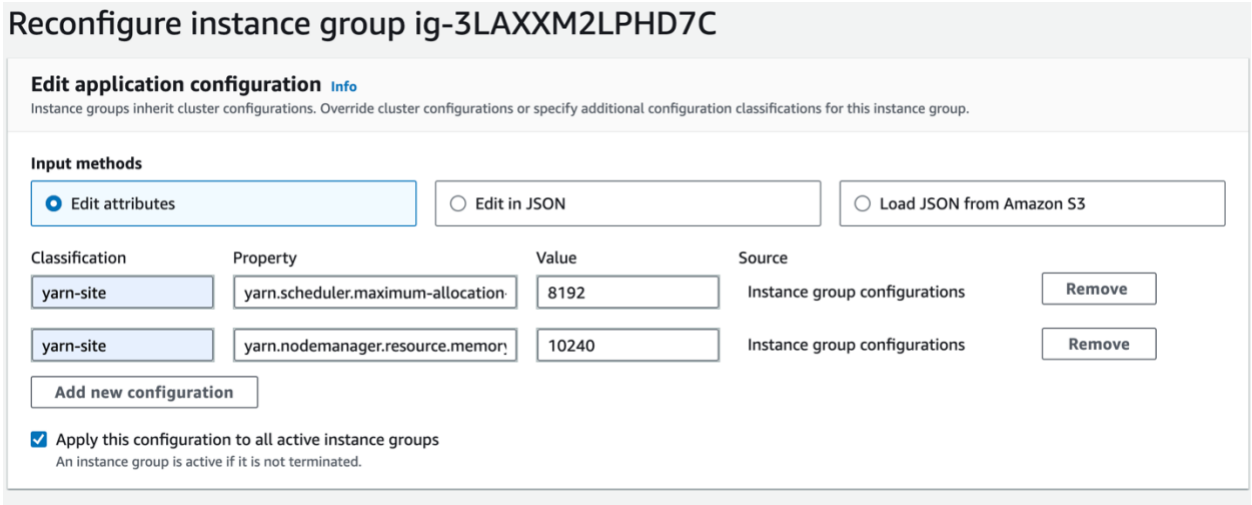
Step 1:

Cluster Setup

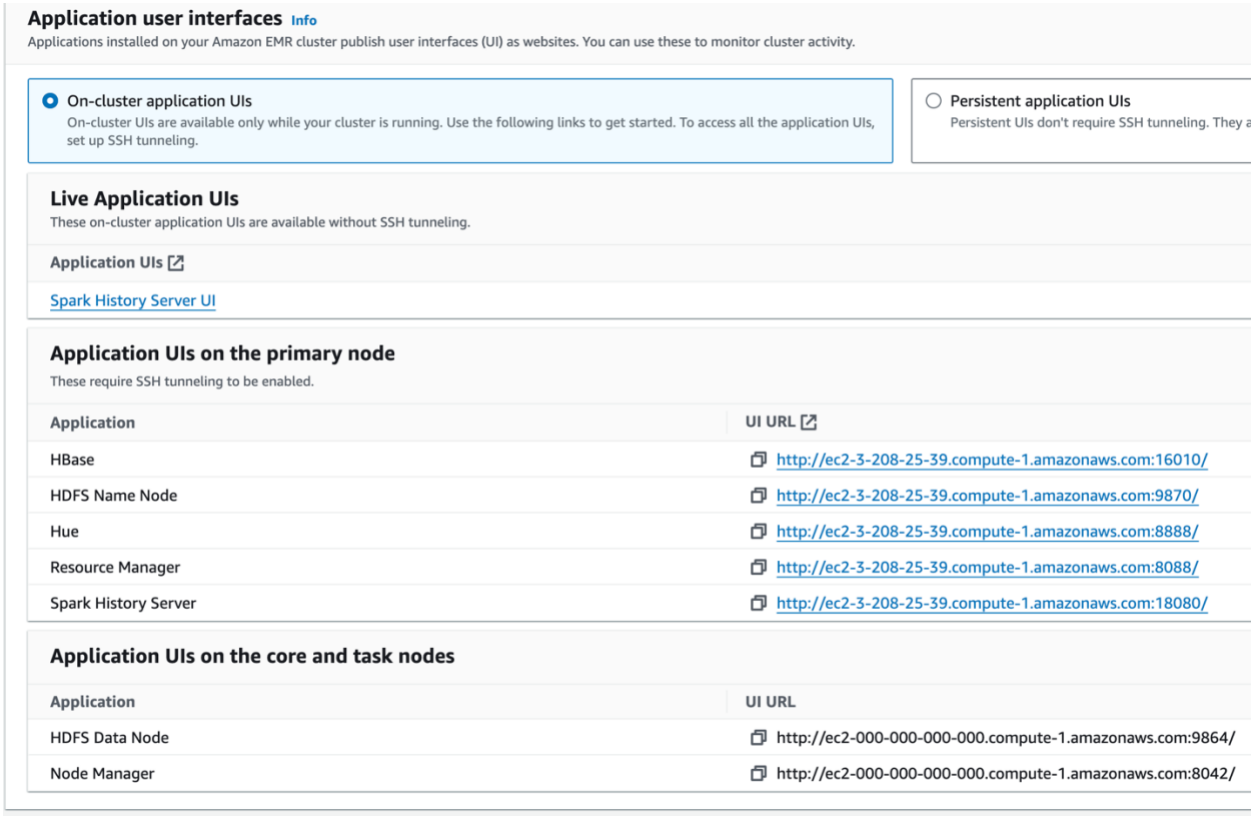
1. Creation using AWS CLI on Canvas

```
eee_W_2806820@runweb111721:~$ aws emr create-cluster   --name "UPGRADLUSTER"   --release-label emr-6.5.0   --applications Name=Hado
op Name=Hive Name=Hue Name=HBase Name=Spark Name=Sqoop   --use-default-roles     --instance-type m5.xlarge   --instance-count 3   -
-ebs-root-volume-size 20         --visible-to-all-users
{
    "ClusterId": "j-1UXG7KCZ7V7FA",
    "ClusterArn": "arn:aws:elasticmapreduce:us-east-1:263056849140:cluster/j-1UXG7KCZ7V7FA"
}
eee_W_2806820@runweb111721:~$
```

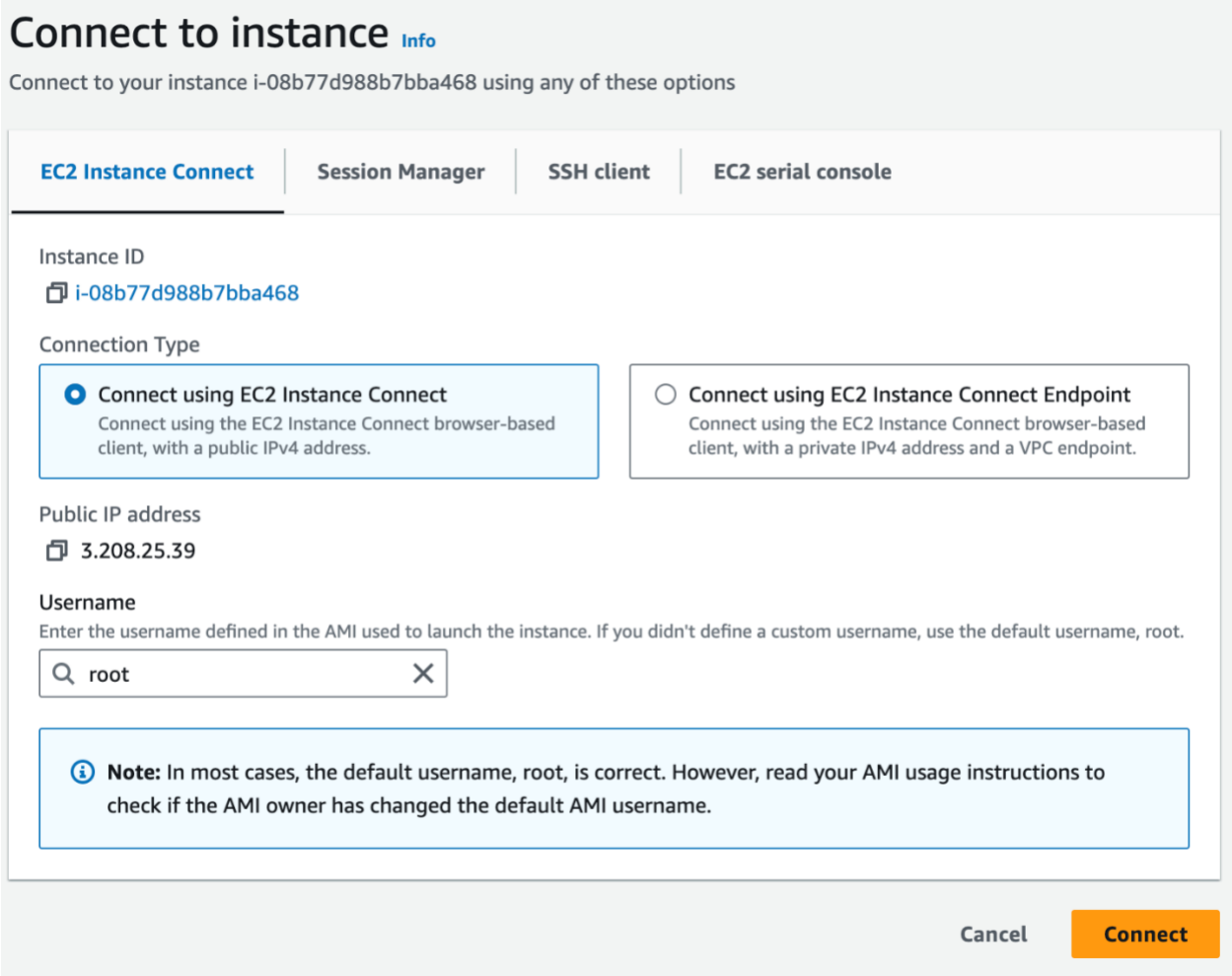
2. Yarn configuration



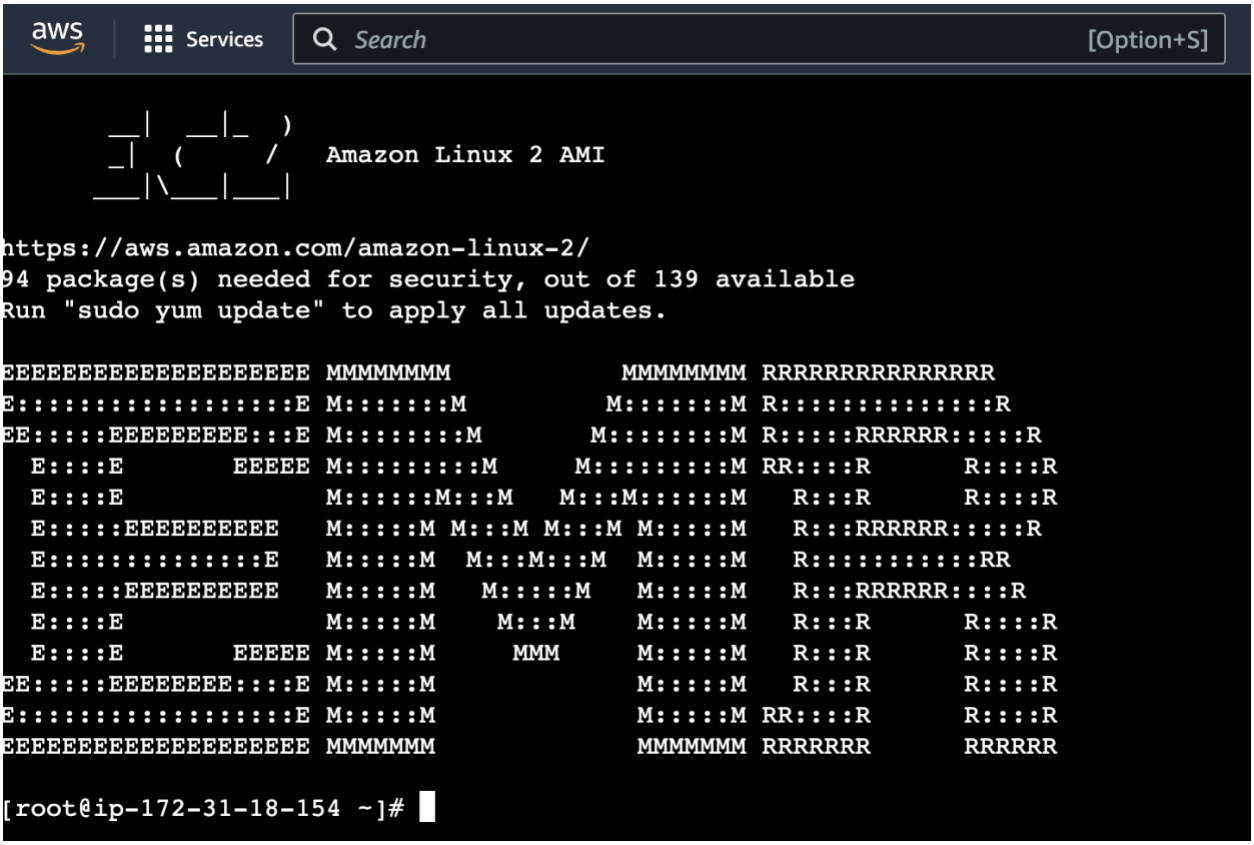
3. Applications installed



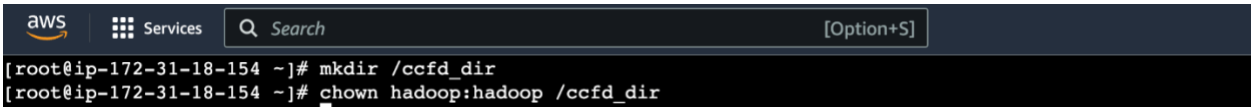
4. SSH using AWS SSM



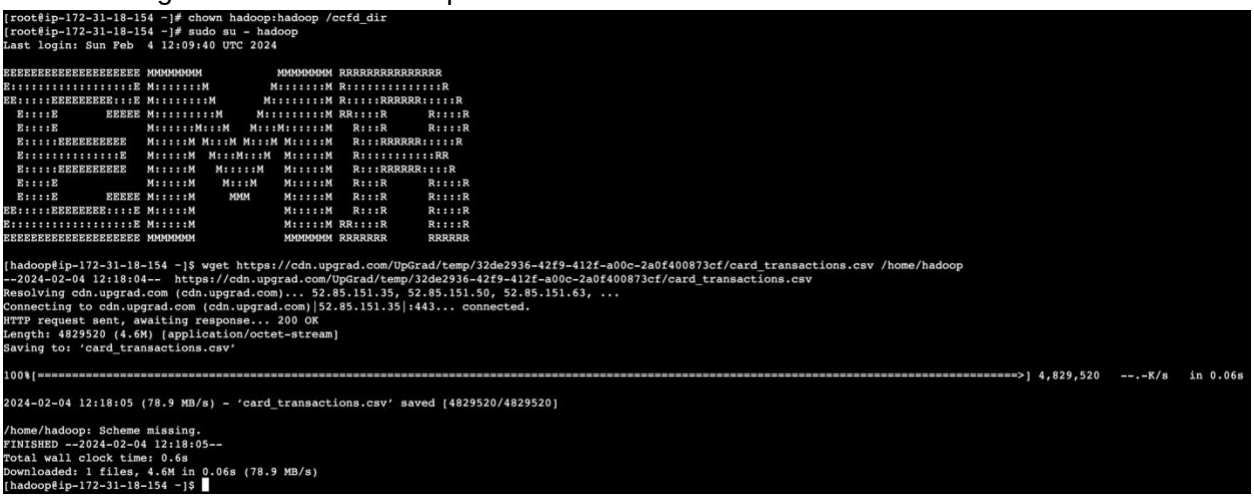
5. Logged in successfully as root user.



6. Making project directory



7. Downloading the csv file to Hadoop home



8. Create Hadoop directory (Ignore SLF4J warning in the below screenshots)

```
[hadoop@ip-172-31-18-154 ~]$ hadoop fs -mkdir /ccfd_dir
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/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.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]
[hadoop@ip-172-31-18-154 ~]$ hadoop fs -mkdir /ccfd_dir/card_transactions
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/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.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]
```

9. Upload csv file to Hadoop

```
aws  Services  Search  [Option+S]
[hadoop@ip-172-31-18-154 ~]$ hadoop fs -put card_transactions.csv /ccfd_dir/card_transactions/
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/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.jar!/org/slf4j/impl/StaticLoggerBinder.cl
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
[hadoop@ip-172-31-18-154 ~]$ hadoop fs ls /ccfd_dir/card_transactions/
```

10. Check csv files in Hadoop directory

```
[hadoop@ip-172-31-18-154 ~]$ hadoop fs -ls /ccfd_dir/card_transactions/
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/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.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]
Found 1 items
-rw-r--r--   1 hadoop hdfsadmingroup    4829520 2024-02-04 12:26 /ccfd_dir/card_transactions/card_transactions.csv
[hadoop@ip-172-31-18-154 ~]$
```

Task 1: Load CSV file to NoSQL database.

1. Login to EMR cluster using ssh and open hive shell(‘hive’)
2. Create a schema/database “ccfd_hive_db”

```
hive> create database IF NOT EXISTS ccfd_hive_db;
OK
Time taken: 1.025 seconds
hive> use ccfd_hive_db;
OK
Time taken: 0.063 seconds
```

3. Set the session parameters appropriately.

```
hive> set hive.auto.convert.join=false;
hive> set hive.stats.autogather=true;
hive> set orc.compress=SNAPPY;
hive> set hive.exec.compress.output=true;
hive> set mapred.output.compression.codec=org.apache.hadoop.io.compress.SnappyCodec; set mapred.output.compression.type=BLOCK;
hive> set mapreduce.map.java.opts=-Xmx5G;
hive> set mapreduce.reduce.java.opts=-Xmx5G;
hive> set mapred.child.java.opts=-Xmx5G -XX:+UseConcMarkSweepGC -XX:-UseGCOverheadLimit;
```

4. Create card_transactions_ext table

```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS CARD_TRANSACTIONS_EXT( `CARD_ID` STRING,
    > `MEMBER_ID` STRING,
    > `AMOUNT` DOUBLE,
    > `POSTCODE` STRING, `POS_ID` STRING, `TRANSACTION_DT` STRING, `STATUS` STRING)
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > LOCATION '/ccfd_dir/card_transactions' TBLPROPERTIES ("skip.header.line.count"="1");
OK
Time taken: 0.069 seconds
hive> select count(1) from CARD_TRANSACTIONS_EXT;
Query ID = hadoop_20240204130805_6a3affc0-78d1-4a87-94a1-66f7ce8d1cc2
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1707048263009_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 3.62 s
--------------------------------------------------------------------------------
OK
53292
Time taken: 4.502 seconds, Fetched: 1 row(s)
```

5. Create ORC table.

```
hive> CREATE
    > TABLE IF NOT EXISTS CARD_TRANSACTIONS_ORC( `CARD_ID` STRING,
    > `MEMBER_ID` STRING,
    > `AMOUNT` DOUBLE,
    > `POSTCODE` STRING, `POS_ID` STRING, `TRANSACTION_DT` TIMESTAMP, `STATUS` STRING)
    > STORED AS ORC
    > TBLPROPERTIES ("orc.compress"="SNAPPY");
OK
Time taken: 0.268 seconds
```

6. Load data into ORC table and converting transaction_dt into timestamp format

```
hive> INSERT OVERWRITE TABLE CARD_TRANSACTIONS_ORC SELECT
    > CARD_ID,
    > MEMBER_ID,
    > AMOUNT,
    > POSTCODE,
    > POS_ID,
    > CAST(FROM_UNIXTIME(UNIX_TIMESTAMP(TRANSACTION_DT, 'dd-MM-yyyy HH:mm:ss')) AS
    > TIMESTAMP),
    >     STATUS
    > FROM CARD_TRANSACTIONS_EXT;
Query ID = hadoop_20240204130859_41620f6c-f22b-484f-9989-df585574f1b9
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1707048263009_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 5.94 s
--------------------------------------------------------------------------------
Loading data to table ccfd_hive_db.card_transactions_orc
OK
Time taken: 7.081 seconds
hive> SELECT count(1) from card_transactions_orc;
OK
53292
```

7. Check year from transaction_dt and verify.

```
hive> SELECT YEAR(transaction_dt), transaction_dt FROM card_transactions_orc LIMIT 10;
OK
2018    2018-02-11 00:00:00
2018    2018-02-11 00:00:00
2018    2018-02-11 00:00:00
2018    2018-02-11 00:00:00
2018    2018-02-11 00:00:00
2018    2018-02-11 00:00:00
2018    2018-02-11 00:00:00
2018    2018-02-11 00:00:00
2018    2018-02-11 00:00:00
Time taken: 2.03 seconds, Fetched: 10 row(s)
```

8. Create hbase integrated table in hive

```
hive> CREATE TABLE CARD_TRANSACTIONS_HBASE( `TRANSACTION_ID` STRING, `CARD_ID` STRING,
    > `MEMBER_ID` STRING,
    > `AMOUNT` DOUBLE, `POSTCODE` STRING, `POS_ID` STRING, `TRANSACTION_DT` TIMESTAMP, `STATUS` STRING)
    > ROW FORMAT DELIMITED
    > STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler' WITH SERDEPROPERTIES (
    > "hbase.columns.mapping"=":key, card_transactions_family:card_id, card_transactions_family:member_id,
    > card_transactions_family:amount, card_transactions_family:postcode, card_transactions_family:pos_id,
    > card_transactions_family:transaction_dt, card_transactions_family:status") TBLPROPERTIES ("hbase.table.name"="card_transactions_hive");
OK
Time taken: 3.98 seconds
```

9. Load data into hbase integrated table and use random UUID as transaction_id(this would be used as row-key)

```
hive> INSERT OVERWRITE TABLE CARD_TRANSACTIONS_HBASE SELECT
    > reflect('java.util.UUID', 'randomUUID') AS TRANSACTION_ID, CARD_ID,
    > MEMBER_ID,
    > AMOUNT,
    >     POSTCODE,
    >     POS_ID,
    >     TRANSACTION_DT,
    >     STATUS
    > FROM CARD_TRANSACTIONS_ORC;
Query ID = hadoop_20240204131839_9f22f827-e4fd-4be4-9983-b453794a8260
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1707048263009_0007)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 6.64 s
--------------------------------------------------------------------------------
OK
Time taken: 78.293 seconds
```

10. Verify some rows from the above created table

```
Time taken: 78.293 seconds
hive> select * from card_transactions_hbase limit 10;
OK
07df2b57-a8d4-4275-9da4-68d129989f53    5527087716468343        632873590564801 8635114.0       41567   649783725699809 2017-12-24 15:46:37     GENUINE
0b3d287b-61eb-4549-9f2b-4031a86ab839    5391723993945313        997128952368160 215255.0        60056   254217670996973 2017-12-13 16:26:15     GENUINE
0e318b80-fe7a-4038-8b3f-ecfee59743be    5155708512920844        300213121454267 3677089.0       11962   698167110333904 2017-03-05 10:45:39     GENUINE
0f52353d-a15c-4505-8a12-399f8ab833a2    6492177672429642        856504383389409 6056381.0       62922   087106851137542 2016-06-04 22:34:42     GENUINE
1ad256be-9a98-4bce-9213-5e66a0ab5b1b    5563292710804718        717127477091175 5301233.0       48816   155123262921334 2016-11-03 15:41:22     GENUINE
20b44db5-4908-4b31-af9c-f20d4d850624    4418227862530505        651536837442483 4761548.0       44436   501310161181891 2017-03-31 01:13:13     GENUINE
2aa77fdc-9de5-4e65-921a-b11ef3fdb0bd    6011920259166638        190431255542160 7193591.0       36032   609723292586452 2017-10-25 23:12:32     GENUINE
2d45adbb-adcc-4014-ac00-abc442ccdb1e    5517690092508277        895174073192784 586893.0        44652   230050674316010 2016-02-27 08:01:56     GENUINE
3244e18d-858a-43ad-ad3c-155152ba6337    3744374493333250 738960224159727 6120469.0       25031   498794716838435 2017-04-04 22:50:43     GENUINE
35af8f2a-7ef8-41f8-a602-4da38ce59542    5555584152987559        671218347984960 4974179.0       14144   036662336806936 2018-01-24 14:23:43     GENUINE
Time taken: 0.206 seconds, Fetched: 10 row(s)
```

11. Check table in hbase

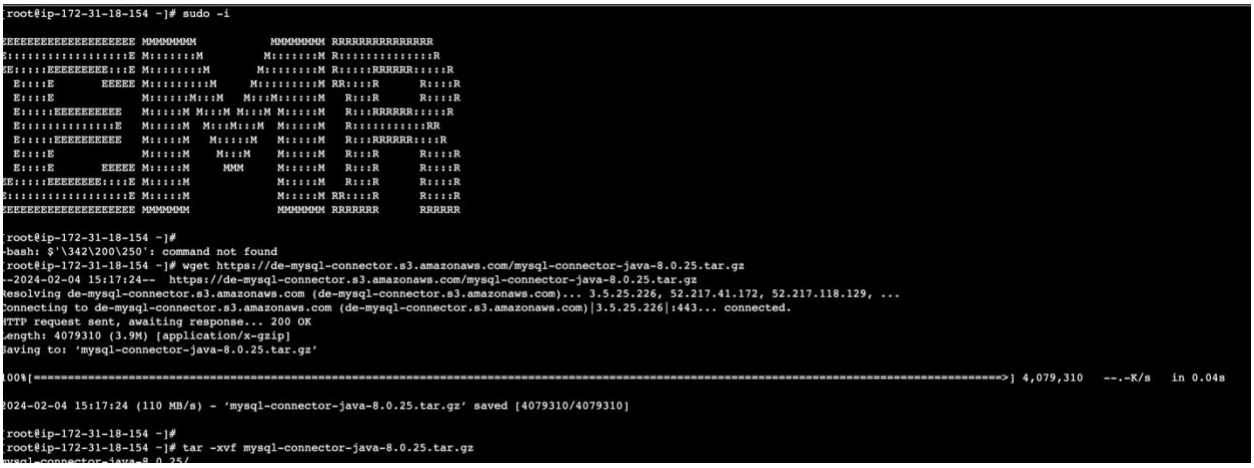
```
[hadoop@ip-172-31-18-154 ~]$ sudo hbase shell
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hbase/lib/client-facing-thirdparty/slf4j-log4j12-1.7.25.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]
HBase Shell
Use "help" to get list of supported commands.
Use "exit" to quit this interactive shell.
For Reference, please visit: http://hbase.apache.org/2.0/book.html#shell
Version 2.4.4-amzn-1, rUnknown, Wed Nov 10 09:50:56 UTC 2021
Took 0.0018 seconds
hbase:001:0> describe 'card_transactions_hive'
Table card_transactions_hive is ENABLED
card_transactions_hive
COLUMN FAMILIES DESCRIPTION
{NAME => 'card_transactions_family', BLOOMFILTER => 'ROW', IN_MEMORY => 'false', VERSIONS => '1', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', COMPRESSION => 'NONE', TT
L => 'FOREVER', MIN_VERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}

1 row(s)
Quota is disabled
Took 0.5141 seconds
hbase:002:0>
```

12. Check count in hbase

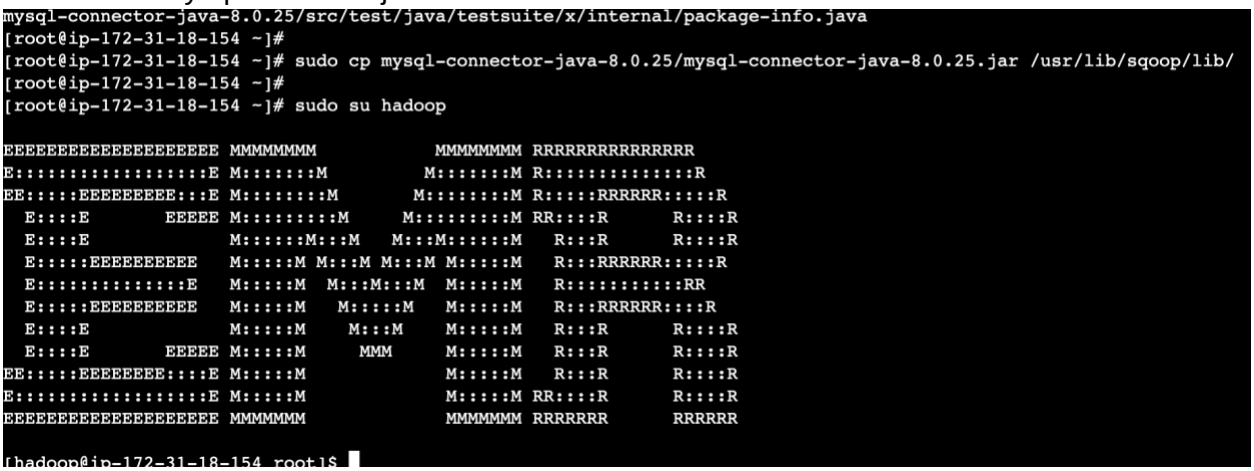
```
hbase:003:0> count 'card_transactions_hive'
Current count: 1000, row: 04d828eb-67ec-4c41-a0de-dc85ae2c170d
Current count: 2000, row: 09abb576-dade-4ac8-9cca-1d218ecc4b64
Current count: 3000, row: 0e965c56-ee8e-43a3-8397-636ad93b27cb
Current count: 4000, row: 131ed467-4a8a-401b-81fb-20ce344e10d3
Current count: 5000, row: 17d15c87-faee-4afd-a0d9-d7595a72444f
Current count: 6000, row: 1cb9809b-cdef-4cb6-99fd-62ec0eb5b34d
Current count: 7000, row: 213e8fda-9fce-4918-92d1-e299346ab1af
Current count: 8000, row: 2603fe6e-f017-42d5-9bb3-06a143626b79
Current count: 9000, row: 2acab997-bfe5-4aef-b5db-4ebe5bb6aa5f
Current count: 10000, row: 2fb2c8cf-3c8b-48f0-9b26-11e2f25bcfe2
Current count: 11000, row: 34cff22d-4d78-4e39-8877-9237905c294a
Current count: 12000, row: 3979ed79-2467-4a72-864b-5ac0a8f5cfde
Current count: 13000, row: 3e37dc3b-411a-4c0b-9650-800a69419488
Current count: 14000, row: 42e7fe0a-c1a7-4a9c-bf5d-a9ec600fb7b2
Current count: 15000, row: 478bcd0a-48aa-43d2-9911-fe531bd9d65c
Current count: 16000, row: 4c528eca-89f6-4ff6-a012-f9a45304c4f9
Current count: 17000, row: 5113eb8a-e771-43f7-b7b4-5057d356c225
Current count: 18000, row: 55c95112-a768-4504-8771-017eb0e63413
Current count: 19000, row: 5a8d2b4a-c54a-4ff2-9314-a7fbbfb5e9de
Current count: 20000, row: 5f7443ba-d424-4f8f-9bb0-170cb54955aa
Current count: 21000, row: 646a9ba4-c62e-4e91-8cce-65e190256d59
Current count: 22000, row: 692312ad-27d1-4498-9cf8-e9c0465a1db5
Current count: 23000, row: 6dc2c825-7fba-4497-a514-1167d45a5573
Current count: 24000, row: 72bfe747-1941-47a4-9aee-2049f5e3a822
Current count: 25000, row: 77bd9505-210d-499d-b196-d8bc6bc45199
Current count: 26000, row: 7c6e5afa-414e-4bab-931b-7e60dfd563d9
Current count: 27000, row: 812f35be-e702-43c3-a6d8-54b362f5a6ba
Current count: 28000, row: 862468cb-69f5-479c-b456-1e97c579c129
Current count: 29000, row: 8acf000e-e08c-40cf-b6cb-6bdc67006b77
Current count: 30000, row: 8fc47b0a-6034-4f47-ab33-d47dc3e605ad
Current count: 31000, row: 94aa94fa-6b5e-49f4-98ca-0e558714d39a
Current count: 32000, row: 995c2d64-a6f4-4cc1-b807-39418c62d8f5
Current count: 33000, row: 9e3bff50-bf35-42d0-a6da-0539a61d3964
Current count: 34000, row: a2e93539-3d82-4b65-97f4-97a9c1e6593b
Current count: 35000, row: a7bdefd0-9260-41dd-b55d-87af641e52a0
Current count: 36000, row: ac60365b-e441-425e-b43b-bd24be171621
Current count: 37000, row: b145cd5a-d141-4c97-b55c-4d5cd6227b6e
Current count: 38000, row: b612378c-0e94-462c-8255-e49fd5b73238
Current count: 39000, row: bb0bf1f5-682c-4b63-95d5-66b903e3232f
Current count: 40000, row: bfc9fc11-b56e-4014-8050-87dc840abaed
Current count: 41000, row: c4ab171d-e83e-4a26-8e98-9e1d2046b228
Current count: 42000, row: c990acdd-1cc3-4f2e-be81-cf3b5e7b9b42
Current count: 43000, row: ce51eef3-6966-46b4-94f1-65e306ce936e
Current count: 44000, row: d32ec724-68b0-4652-8234-021d64e99300
Current count: 45000, row: d80331e3-8967-4d1c-9188-8f03a912ef8b
Current count: 46000, row: dcbda4b6-ecab-413d-9b13-bab46e2641ab
Current count: 47000, row: e1b2d231-9c00-4915-a44c-61150c11feb2
Current count: 48000, row: e6cb4653-4674-4802-9f9e-e9c416a80db2
Current count: 49000, row: ebcc2cd6-4890-46bf-ba29-5dfa38c7c84c
Current count: 50000, row: f07b7ecd-405f-409d-b5eb-7c4aa63d6d23
Current count: 51000, row: f5218255-cf02-4658-ace6-1a74d34ce3ad
Current count: 52000, row: f9d9b9a8-10ad-4088-bda4-a8c01571465d
Current count: 53000, row: fe77ddf2-7ded-4a11-b843-8d58260a64b6
53292 row(s)
Took 1.5885 seconds
=> 53292
hbase:004:0>
```

Task 2: Ingest Data from RDS to Hadoop

1. Setup permissions



2. Add mysql connector jar



3. Sqoop commands start for importing card_member

```
[hadoop@ip-172-31-18-154 root]$ sqoop import --connect jdbc:mysql://upgradawsrdsl.cyaielc9bmnf.us-east-1.rds.amazonaws.com/cred_financials_data \
> --username upgraduser \
> --password upgraduser \
> --table card_member \
> --null-string 'NA' \
> --null-non-string '\\N' \
> --delete-target-dir \
> --target-dir '/ccfd_dir/card_member' \
> -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.
find: failed to restore initial working directory: Permission denied
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/share/aws/redshift/jdbc/redshift-jdbc42-1.2.37.1061.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hbase/lib/client-facing-thirdparty/slf4j-log4j12-1.7.25.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]
2024-02-04 15:56:33,250 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
2024-02-04 15:56:33,365 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
2024-02-04 15:56:33,473 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
2024-02-04 15:56:33,473 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 t
he driver class is generally unnecessary.
2024-02-04 15:56:33,984 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `card_member` AS t LIMIT 1
2024-02-04 15:56:34,025 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `card_member` AS t LIMIT 1
2024-02-04 15:56:34,048 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
2024-02-04 15:56:36,851 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hadoop/compile/0bef32bb10011daad95191f42bc43300/card_member.jar
2024-02-04 15:56:37,893 INFO tool.ImportTool: Destination directory /ccfd_dir/card_member is not present, hence not deleting.
2024-02-04 15:56:37,893 WARN manager.MySQLManager: It looks like you are importing from mysql.
2024-02-04 15:56:37,893 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
2024-02-04 15:56:37,893 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
2024-02-04 15:56:37,894 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
2024-02-04 15:56:37,904 INFO mapreduce.ImportJobBase: Beginning import of card_member
2024-02-04 15:56:37,911 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
2024-02-04 15:56:37,927 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
2024-02-04 15:56:38,182 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-18-154.ec2.internal/172.31.18.154:8032
2024-02-04 15:56:38,381 INFO client.AHSProxy: Connecting to Application History server at ip-172-31-18-154.ec2.internal/172.31.18.154:10200
2024-02-04 15:56:38,802 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1707048263009_0031
```

```
2024-02-04 15:56:38,802 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1707048263009_0031
2024-02-04 15:56:39,684 INFO db.DBInputFormat: Using read commited transaction isolation
2024-02-04 15:56:39,736 INFO mapreduce.JobSubmitter: number of splits:1
2024-02-04 15:56:39,981 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1707048263009_0031
2024-02-04 15:56:39,983 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-02-04 15:56:40,180 INFO conf.Configuration: resource-types.xml not found
2024-02-04 15:56:40,182 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-02-04 15:56:40,267 INFO impl.YarnClientImpl: Submitted application application_1707048263009_0031
2024-02-04 15:56:40,300 INFO mapreduce.Job: The url to track the job: http://ip-172-31-18-154.ec2.internal:20888/proxy/application_1707048263009_0031/
2024-02-04 15:56:40,300 INFO mapreduce.Job: Running job: job_1707048263009_0031
2024-02-04 15:56:46,352 INFO mapreduce.Job: Job job_1707048263009_0031 running in uber mode : false
2024-02-04 15:56:46,353 INFO mapreduce.Job:  map 0% reduce 0%
2024-02-04 15:56:51,395 INFO mapreduce.Job:  map 100% reduce 0%
2024-02-04 15:56:51,402 INFO mapreduce.Job: Job job_1707048263009_0031 completed successfully
2024-02-04 15:56:51,486 INFO mapreduce.Job: Counters: 33
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=247567
                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=85081
                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)=267648
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=2788
                Total vcore-milliseconds taken by all map tasks=2788
                Total megabyte-milliseconds taken by all map tasks=8564736
        Map-Reduce Framework
                Map input records=999
                Map output records=999
                Input split bytes=85
                Spilled Records=0
                Failed Shuffles=0
```

```
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=247567
                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=85081
                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)=267648
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=2788
                Total vcore-milliseconds taken by all map tasks=2788
                Total megabyte-milliseconds taken by all map tasks=8564736
        Map-Reduce Framework
                Map input records=999
                Map output records=999
                Input split bytes=85
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=70
                CPU time spent (ms)=1500
                Physical memory (bytes) snapshot=307757056
                Virtual memory (bytes) snapshot=4403085312
                Total committed heap usage (bytes)=256376832
                Peak Map Physical memory (bytes)=307757056
                Peak Map Virtual memory (bytes)=4403085312
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=85081
2024-02-04 15:56:51,491 INFO mapreduce.ImportJobBase: Transferred 83.0869 KB in 13.5531 seconds (6.1305 KB/sec)
2024-02-04 15:56:51,493 INFO mapreduce.ImportJobBase: Retrieved 999 records.
[hadoop@ip-172-31-18-154 root]$
```

Total 999 records are imported

4. Sqoop commands imports for Member_score data

```
[hadoop@ip-172-31-18-154 root]$ sqoop import --connect jdbc:mysql://upgradawsrds1.cyaielc9bmnf.us-east-1.rds.amazonaws.com/cred_financials_data \
> --username upgraduser \
> --password upgraduser \
> --table member_score \
> --null-string 'NA' \
> --null-non-string '\\N' \
> --delete-target-dir \
> --target-dir '/ccfd_dir/member_score' \
> -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.
find: failed to restore initial working directory: Permission denied
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/share/aws/redshift/jdbc/redshift-jdbc42-1.2.37.1061.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hbase/lib/client-facing-thirdparty/slf4j-log4j12-1.7.25.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]
2024-02-04 15:58:28,527 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
2024-02-04 15:58:28,605 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
2024-02-04 15:58:28,707 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
2024-02-04 15:58:28,707 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 t
he driver class is generally unnecessary.
2024-02-04 15:58:29,165 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `member_score` AS t LIMIT 1
2024-02-04 15:58:29,206 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `member_score` AS t LIMIT 1
2024-02-04 15:58:29,228 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
2024-02-04 15:58:31,505 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hadoop/compile/a036d96e4ea0155877542947e704e405/member_score.jar
2024-02-04 15:58:32,941 INFO tool.ImportTool: Destination directory /ccfd_dir/member_score is not present, hence not deleting.
2024-02-04 15:58:32,942 WARN manager.MySQLManager: It looks like you are importing from mysql.
2024-02-04 15:58:32,942 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
2024-02-04 15:58:32,942 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
2024-02-04 15:58:32,942 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
2024-02-04 15:58:32,953 INFO mapreduce.ImportJobBase: Beginning import of member_score
2024-02-04 15:58:32,959 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
2024-02-04 15:58:32,977 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
2024-02-04 15:58:33,259 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-18-154.ec2.internal/172.31.18.154:8032
2024-02-04 15:58:33,514 INFO client.AHSProxy: Connecting to Application History server at ip-172-31-18-154.ec2.internal/172.31.18.154:10200
2024-02-04 15:58:34,026 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1707048263009_0032
```

```
2024-02-04 15:58:34,026 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1707048263009_0032
2024-02-04 15:58:34,863 INFO db.DBInputFormat: Using read commited transaction isolation
2024-02-04 15:58:34,915 INFO mapreduce.JobSubmitter: number of splits:1
2024-02-04 15:58:35,171 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1707048263009_0032
2024-02-04 15:58:35,172 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-02-04 15:58:35,342 INFO conf.Configuration: resource-types.xml not found
2024-02-04 15:58:35,343 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-02-04 15:58:35,412 INFO impl.YarnClientImpl: Submitted application application_1707048263009_0032
2024-02-04 15:58:35,456 INFO mapreduce.Job: The url to track the job: http://ip-172-31-18-154.ec2.internal:20888/proxy/application_1707048263009_0032/
2024-02-04 15:58:35,457 INFO mapreduce.Job: Running job: job_1707048263009_0032
2024-02-04 15:58:41,511 INFO mapreduce.Job: Job job_1707048263009_0032 running in uber mode : false
2024-02-04 15:58:41,511 INFO mapreduce.Job:  map 0% reduce 0%
2024-02-04 15:58:46,560 INFO mapreduce.Job:  map 100% reduce 0%
2024-02-04 15:58:46,567 INFO mapreduce.Job: Job job_1707048263009_0032 completed successfully
2024-02-04 15:58:46,673 INFO mapreduce.Job: Counters: 33
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=247514
                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=19980
                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)=270432
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=2817
                Total vcore-milliseconds taken by all map tasks=2817
                Total megabyte-milliseconds taken by all map tasks=8653824
        Map-Reduce Framework
                Map input records=999
                Map output records=999
                Input split bytes=85
                Spilled Records=0
                Failed Shuffles=0
```

```
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=247514
                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=19980
                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)=270432
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=2817
                Total vcore-milliseconds taken by all map tasks=2817
                Total megabyte-milliseconds taken by all map tasks=8653824
        Map-Reduce Framework
                Map input records=999
                Map output records=999
                Input split bytes=85
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=79
                CPU time spent (ms)=1400
                Physical memory (bytes) snapshot=310341632
                Virtual memory (bytes) snapshot=4397051904
                Total committed heap usage (bytes)=330825728
                Peak Map Physical memory (bytes)=310341632
                Peak Map Virtual memory (bytes)=4397051904
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=19980
2024-02-04 15:58:46,679 INFO mapreduce.ImportJobBase: Transferred 19.5117 KB in 13.6907 seconds (1.4252 KB/sec)
2024-02-04 15:58:46,682 INFO mapreduce.ImportJobBase: Retrieved 999 records.
[hadoop@ip-172-31-18-154 root]$
```

Total 999 records are imported

5. Create external table member_score_ext and card_member_ext in hive that points to data imported form sqoop command

```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS member_score_ext (
    >   member_id STRING,
    >   score INT
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > LOCATION '/ccfd_dir/member_score';
OK
Time taken: 0.074 seconds
hive> SELECT * FROM member_score_ext LIMIT 10;
OK
000037495066290 339
000117826301530 289
001147922084344 393
001314074991813 225
001739553947511 642
003761426295463 413
004494068832701 217
006836124210484 504
006991872634058 697
007955566230397 372
Time taken: 0.134 seconds, Fetched: 10 row(s)
```

```
Hive Session ID = 149e0c97-1adf-4de4-8196-ef4cb456c74d
hive>
    >
    >
    > CREATE EXTERNAL TABLE IF NOT EXISTS card_member_ext (
    >   card_id STRING,
    >   member_id STRING,
    >   member_joining_dt TIMESTAMP,
    >   card_purchase_dt STRING,
    >   country STRING,
    >   city STRING
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > LOCATION '/ccfd_dir/card_member';
OK
Time taken: 0.754 seconds
```

6. Create ORC table for card_member to save space and load it using card_member_ext

```
hive>
    > CREATE TABLE IF NOT EXISTS CARD_MEMBER_ORC(
    >     `CARD_ID` STRING,
    >     `MEMBER_ID` STRING,
    >     `MEMBER_JOINING_DT` TIMESTAMP,
    >     `CARD_PURCHASE_DT` STRING,
    >     `COUNTRY` STRING,
    >     `CITY` STRING
    > )
    > STORED AS ORC
    > TBLPROPERTIES ("orc.compress"="SNAPPY");
OK
Time taken: 0.806 seconds
hive> INSERT OVERWRITE TABLE CARD_MEMBER_ORC
    > SELECT
    >     CARD_ID,
    >     MEMBER_ID,
    >     MEMBER_JOINING_DT,
    >     CARD_PURCHASE_DT,
    >     COUNTRY,
    >     CITY
    > FROM
    >     CARD_MEMBER_EXT;
Query ID = hadoop_20240204173045_a612fc73-7ade-4e69-92a9-7cdd62d4bf07
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1707048263009_0036)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 4.50 s
--------------------------------------------------------------------------------
Loading data to table default.card_member_orc
OK
Time taken: 9.118 seconds
hive> select count(1) from CARD_MEMBER_ORC;
OK
999
Time taken: 0.483 seconds, Fetched: 1 row(s)
```

7. Create ORC format table for member_score and load it using member_score_ext

```
hive> CREATE TABLE IF NOT EXISTS MEMBER_SCORE_ORC(
    >     `MEMBER_ID` STRING,
    > `SCORE` INT )
    > STORED AS ORC
    > TBLPROPERTIES ("orc.compress"="SNAPPY");
OK
Time taken: 0.051 seconds
hive> INSERT OVERWRITE TABLE MEMBER_SCORE_ORC
    > SELECT
    > MEMBER_ID,
    > SCORE FROM
    >     MEMBER_SCORE_EXT;
Query ID = hadoop_20240204173317_7231eaf3-6ec5-4ceb-abc6-f0ac07479f27
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1707048263009_0036)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 3.76 s
----------------------------------------------------------------------------------------------
Loading data to table default.member_score_orc
OK
Time taken: 4.791 seconds
hive> select count(1) from MEMBER_SCORE_ORC;
OK
999
Time taken: 0.137 seconds, Fetched: 1 row(s)
```

8. Check some sample rows from card_member_orc and member_score_orc

```
Time taken: 0.137 seconds, Fetched: 1 row(s)
hive> SELECT * FROM CARD_MEMBER_ORC LIMIT 10;
OK
340028465709212 009250698176266 2012-02-08 06:04:13     05/13   United States   Barberton
340054675199675 835873341185231 2017-03-10 09:24:44     03/17   United States   Fort Dodge
340082915339645 512969555857346 2014-02-15 06:30:30     07/14   United States   Graham
340134186926007 887711945571282 2012-02-05 01:21:58     02/13   United States   Dix Hills
340265728490548 680324265406190 2014-03-29 07:49:14     11/14   United States   Rancho Cucamonga
340268219434811 929799084911715 2012-07-08 02:46:08     08/12   United States   San Francisco
340379737226464 089615510858348 2010-03-10 00:06:42     09/10   United States   Clinton
340383645652108 181180599313885 2012-02-24 05:32:44     10/16   United States   West New York
340803866934451 417664728506297 2015-05-21 04:30:45     08/17   United States   Beaverton
340889618969736 459292914761635 2013-04-23 08:40:11     11/15   United States   West Palm Beach
Time taken: 0.131 seconds, Fetched: 10 row(s)
```

```
hive> SELECT * FROM MEMBER_SCORE_ORC LIMIT 10;
OK
000037495066290 339
000117826301530 289
001147922084344 393
001314074991813 225
001739553947511 642
003761426295463 413
004494068832701 217
006836124210484 504
006991872634058 697
007955566230397 372
Time taken: 0.117 seconds, Fetched: 10 row(s)
```

Task 3: Lookup table creation

1. Setup permissions for hbase directory

```
[hadoop@ip-172-31-18-154 root]$ sudo chown hadoop -R /var/log/hbase
[hadoop@ip-172-31-18-154 root]$ mkdir /var/log/hbase/user/
[hadoop@ip-172-31-18-154 root]$ mkdir /var/log/hbase/user/hadoop
```

2. Create and hbase-hive integrated table in hive, with underlying storage in hbase

```
hive> CREATE TABLE LOOKUP_DATA_HBASE(CARD_ID STRING,UCL DOUBLE, SCORE INT, POSTCODE STRING, TRANSACTION_DT TIMESTAMP) STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler' WITH S
ERDEPROPERTIES ("hbase.columns.mapping"=":key, lookup_card_family:ucl, lookup_card_family:score, lookup_transaction_family:postcode, lookup_transaction_family:transaction_dt") TBLPROPER
TIES ("hbase.table.name" = "lookup_data_hive");
OK
Time taken: 2.028 seconds
```

3. Check table details in hbase

```
hbase:001:0> describe 'lookup_data_hive'
Table lookup_data_hive is ENABLED
lookup_data_hive
COLUMN FAMILIES DESCRIPTION
{NAME => 'lookup_card_family', BLOOMFILTER => 'ROW', IN_MEMORY => 'false', VERSIONS => '1', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', COMPRESSION => 'NONE', TTL => '
FOREVER', MIN_VERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}

{NAME => 'lookup_transaction_family', BLOOMFILTER => 'ROW', IN_MEMORY => 'false', VERSIONS => '1', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', COMPRESSION => 'NONE', T
TL => 'FOREVER', MIN_VERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}

2 row(s)
Quota is disabled
Took 0.5475 seconds
```

Task 4: Loading the data in lookup table

1. Create table ranked_card_transactions_orc and card_ucl_orc

```
hive> CREATE TABLE IF NOT EXISTS RANKED_CARD_TRANSACTIONS_ORC( `CARD_ID` STRING,
    > `AMOUNT` DOUBLE,
    > `POSTCODE` STRING,
    > `TRANSACTION_DT` TIMESTAMP,
    > `RANK` INT)
    > STORED AS ORC
    > TBLPROPERTIES ("orc.compress"="SNAPPY");
OK
Time taken: 0.085 seconds
hive> CREATE TABLE IF NOT EXISTS CARD_UCL_ORC(
    > `CARD_ID` STRING,
    > `UCL` DOUBLE)
    > STORED AS ORC
    > TBLPROPERTIES ("orc.compress"="SNAPPY");
OK
Time taken: 0.046 seconds
```

2. Insert rows into the above 2 tables

```
hive> INSERT OVERWRITE TABLE RANKED_CARD_TRANSACTIONS_ORC SELECT
    >     B.CARD_ID,
    >     B.AMOUNT,
    >     B.POSTCODE,
    >     B.TRANSACTION_DT,
    >     B.RANK
    > FROM ( SELECT
    >         A.CARD_ID,
    >         A.AMOUNT,
    >         A.POSTCODE,
    >         A.TRANSACTION_DT,
    >         RANK() OVER (
    >             PARTITION BY A.CARD_ID
    >             ORDER BY A.TRANSACTION_DT DESC, A.AMOUNT DESC
    >         ) AS RANK
    > FROM ( SELECT
    >             CARD_ID,
    >             AMOUNT,
    >             POSTCODE,
    >             TRANSACTION_DT
    >         FROM CARD_TRANSACTIONS_ORC
    > WHERE STATUS = 'GENUINE' )A
    > )B
    > WHERE B.RANK <= 10;
Query ID = hadoop_20240204190319_eb423466-ff82-469e-88a2-a62e9b97a3cd
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1707048263009_0054)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 5.74 s
--------------------------------------------------------------------------------
Loading data to table ccfd_hive_db.ranked_card_transactions_orc
OK
Time taken: 7.144 seconds
```

```
hive> INSERT OVERWRITE TABLE CARD_UCL_ORC
    > SELECT
    > A.CARD_ID,
    >     (A.AVERAGE + (3 * A.STANDARD_DEVIATION)) AS UCL
    > FROM (
    >     SELECT
    >         CARD_ID,
    >         AVG(AMOUNT) AS AVERAGE,
    >         STDDEV(AMOUNT) AS STANDARD_DEVIATION
    >     FROM RANKED_CARD_TRANSACTIONS_ORC
    >     GROUP BY CARD_ID
    > ) A;
Query ID = hadoop_20240204190419_eab0c6d5-eb9a-4c5a-8277-10d56d86b340
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1707048263009_0054)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 4.89 s
--------------------------------------------------------------------------------
Loading data to table ccfd_hive_db.card_ucl_orc
OK
Time taken: 5.994 seconds
```

3. Load data in Lookup table integrated with hbase

```
hive> INSERT OVERWRITE TABLE LOOKUP_DATA_HBASE
    > SELECT
    >     RCTO.CARD_ID,
    >     CUO.UCL,
    >     CMS.SCORE,
    >     RCTO.POSTCODE,
    >     RCTO.TRANSACTION_DT
    > FROM RANKED_CARD_TRANSACTIONS_ORC RCTO
    > JOIN CARD_UCL_ORC CUO
    >     ON CUO.CARD_ID = RCTO.CARD_ID
    > JOIN (
    >     SELECT DISTINCT
    >         CARD.CARD_ID,
    >         SCORE.SCORE
    >     FROM CARD_MEMBER_ORC CARD
    >     JOIN MEMBER_SCORE_ORC SCORE
    >         ON CARD.MEMBER_ID = SCORE.MEMBER_ID
    > ) AS CMS
    >     ON RCTO.CARD_ID = CMS.CARD_ID
    > WHERE RCTO.RANK = 1;
Query ID = hadoop_20240204190511_3be76282-dfdf-4827-8d40-c4a42c87b7e2
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1707048263009_0054)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 5 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 3 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 4 ...... container     SUCCEEDED      2          2        0        0       0       0
Map 2 .......... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 05/05  [==========================>>] 100%  ELAPSED TIME: 9.62 s
--------------------------------------------------------------------------------
OK
Time taken: 13.873 seconds
```

4. Look at sample rows in lookup table

```
hive> select * from lookup_data_hbase limit 10;
OK
340028465709212 1.6331555548882347E7    233     24658   2018-01-02 03:25:35
340054675199675 1.4156079786189131E7    631     50140   2018-01-15 19:43:23
340082915339645 1.5285685330791477E7    407     17844   2018-01-26 19:03:47
340134186926007 1.5239767522438552E7    614     67576   2018-01-18 23:12:50
340265728490548 1.6084916712555619E7    202     72435   2018-01-21 02:07:35
340268219434811 1.2507323937605347E7    415     62513   2018-01-16 04:30:05
340379737226464 1.4198310998368105E7    229     26656   2018-01-27 00:19:47
340383645652108 1.4091750460468251E7    645     34734   2018-01-29 01:29:12
340803866934451 1.0843341196185412E7    502     87525   2018-01-31 04:23:57
340889618969736 1.3217942365515321E7    330     61341   2018-01-31 21:57:18
Time taken: 0.156 seconds, Fetched: 10 row(s)
hive> select count(*) from lookup_data_hbase;
Query ID = hadoop_20240204190624_8d167f8d-76d8-43fb-9a81-26dcb8f517f0
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1707048263009_0054)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 5.71 s
----------------------------------------------------------------------------------------------
OK
999
```

5. Count the total rows in lookup table in hbase

```
hbase:001:0> count 'lookup_data_hive'
999 row(s)
Took 1.1397 seconds
```

6. Look at rows in hbase table

```
hbase:001:0> scan 'lookup_data_hive'
ROW                                     COLUMN+CELL
 340028465709212                        column=lookup_card_family:score, timestamp=2024-02-04T19:05:25.145, value=233
 340028465709212                        column=lookup_card_family:ucl, timestamp=2024-02-04T19:05:25.145, value=1.6331555548882347E7
 340028465709212                        column=lookup_transaction_family:postcode, timestamp=2024-02-04T19:05:25.145, value=24658
 340028465709212                        column=lookup_transaction_family:transaction_dt, timestamp=2024-02-04T19:05:25.145, value=2018-01-02 03:25:35
 340054675199675                        column=lookup_card_family:score, timestamp=2024-02-04T19:05:25.145, value=631
 340054675199675                        column=lookup_card_family:ucl, timestamp=2024-02-04T19:05:25.145, value=1.4156079786189131E7
 340054675199675                        column=lookup_transaction_family:postcode, timestamp=2024-02-04T19:05:25.145, value=50140
 340054675199675                        column=lookup_transaction_family:transaction_dt, timestamp=2024-02-04T19:05:25.145, value=2018-01-15 19:43:23
 340082915339645                        column=lookup_card_family:score, timestamp=2024-02-04T19:05:25.145, value=407
 340082915339645                        column=lookup_card_family:ucl, timestamp=2024-02-04T19:05:25.145, value=1.5285685330791477E7
 340082915339645                        column=lookup_transaction_family:postcode, timestamp=2024-02-04T19:05:25.145, value=17844
 340082915339645                        column=lookup_transaction_family:transaction_dt, timestamp=2024-02-04T19:05:25.145, value=2018-01-26 19:03:47
 340134186926007                        column=lookup_card_family:score, timestamp=2024-02-04T19:05:25.145, value=614
 340134186926007                        column=lookup_card_family:ucl, timestamp=2024-02-04T19:05:25.145, value=1.5239767522438552E7
 340134186926007                        column=lookup_transaction_family:postcode, timestamp=2024-02-04T19:05:25.145, value=67576
 340134186926007                        column=lookup_transaction_family:transaction_dt, timestamp=2024-02-04T19:05:25.145, value=2018-01-18 23:12:50
 340265728490548                        column=lookup_card_family:score, timestamp=2024-02-04T19:05:25.145, value=202
 340265728490548                        column=lookup_card_family:ucl, timestamp=2024-02-04T19:05:25.145, value=1.6084916712555619E7
 340265728490548                        column=lookup_transaction_family:postcode, timestamp=2024-02-04T19:05:25.145, value=72435
 340265728490548                        column=lookup_transaction_family:transaction_dt, timestamp=2024-02-04T19:05:25.145, value=2018-01-21 02:07:35
```