Scripts Execution

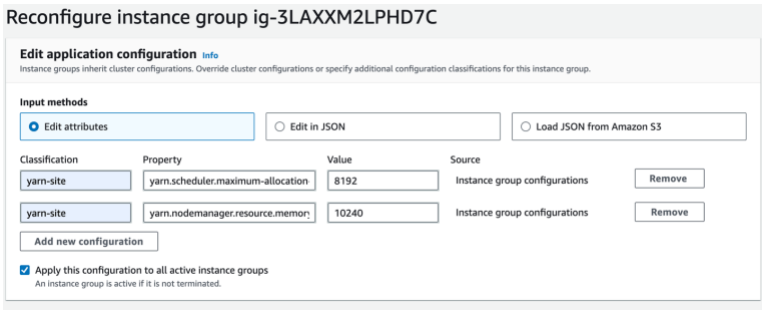
Screenshots of the execution of the scripts written:

Cluster Setup

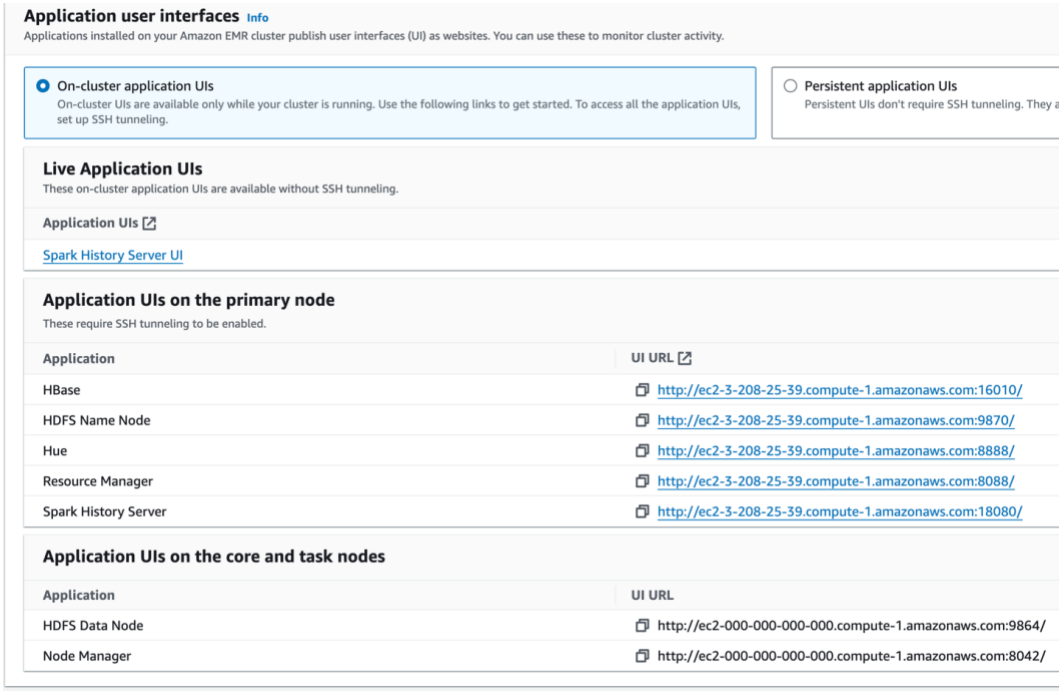
1. Creation using AWS CLI on Canvas

```
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:~$
```

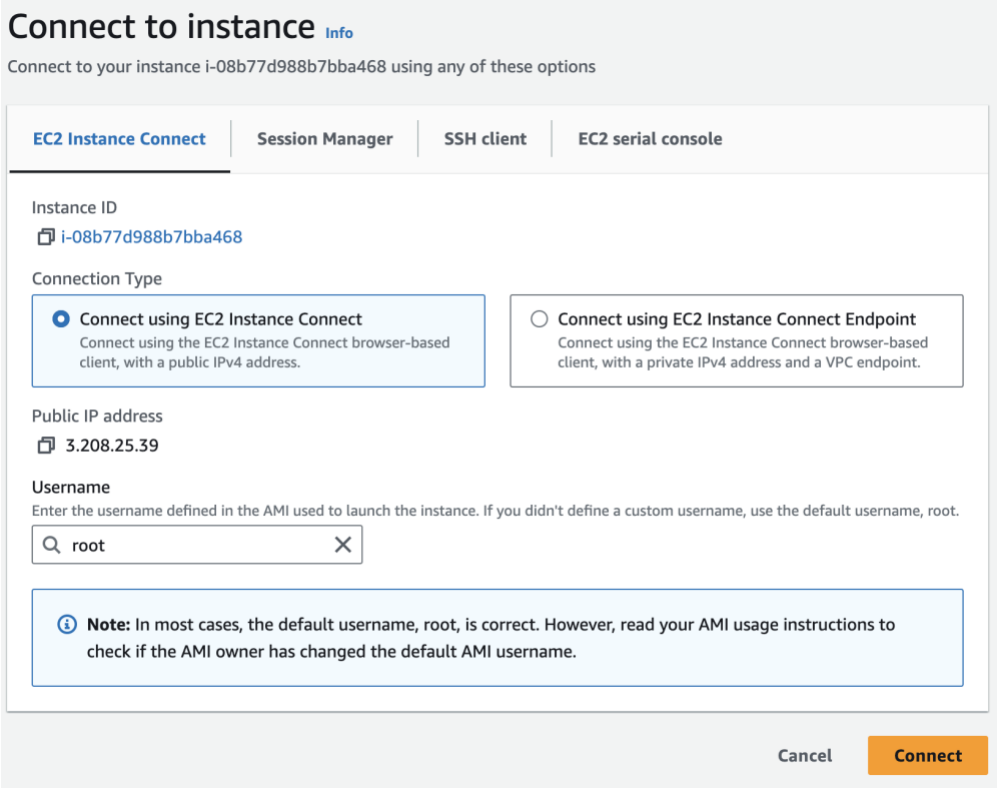
2. Yarn configuration

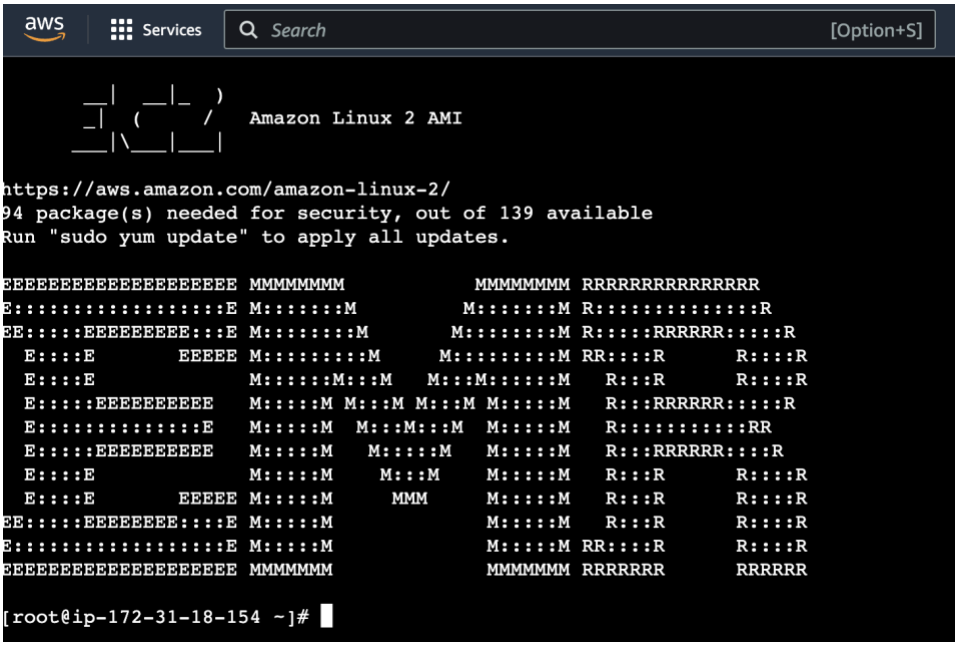


3. Applications UI's

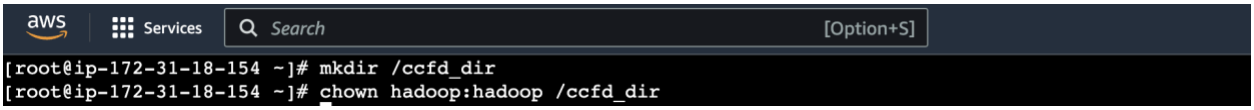


4. Connecting to EMR cluster (SSH using AWS SSM)

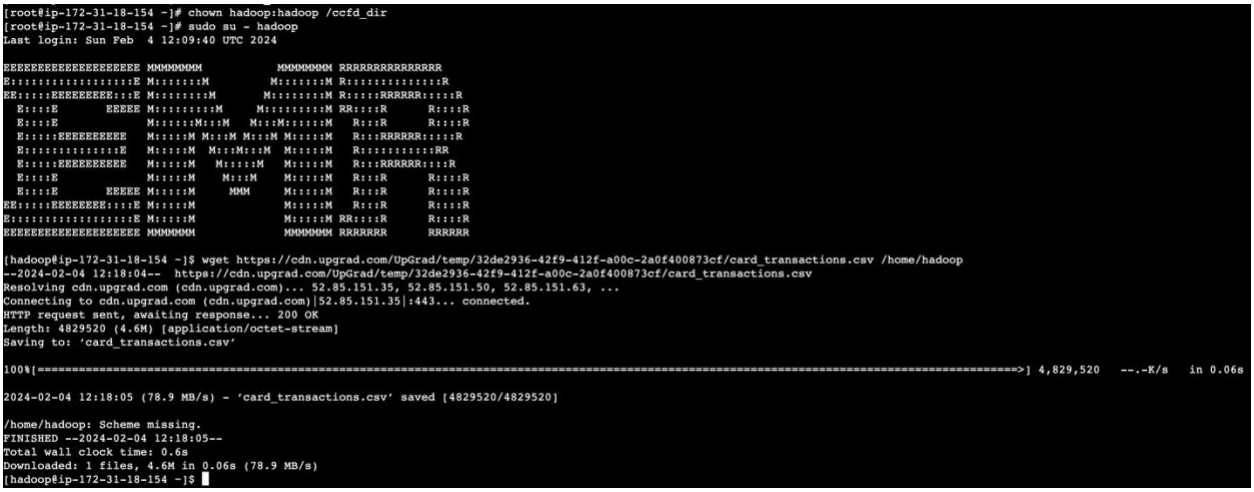


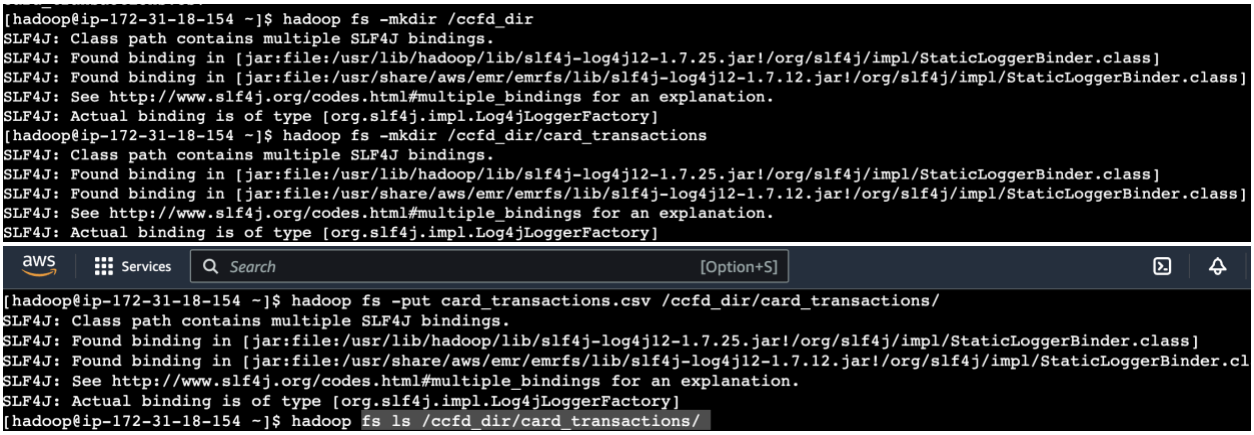


5. Switching to root user and making the project directory named ‘ccdf_dir’



6. Downloading the card_transactions.csv file and transferring it into ‘ccdf_dir’ in hdfs





Task 1: Load the transactions history data (card_transactions.csv) in a NoSQL database

1. Creating a database named ccfd_hive_db and using it for further operations

```
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
```

2. Setting up a Hive session to enhance performance and efficiently manage resource utilisation

```
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;
```

3. Creating an external table named ‘CARD_TRANSACTIONS_EXT’ under same database and loading the data in csv file in HDFS

```
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)
```

4. Creating ‘CARD_TRANSACTIONS_ORC’ table using the ORC format

```
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
```

5. Inserting data into ‘CARD_TRANSACTIONS_ORC’ table from ‘CARD_TRANSACTIONS_EXT' and also converting TRANSACTION_DT into unix 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
```

6. Next, we extracted the Year information from the "transaction_dt" column to validate it.

```
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
2018    2018-02-11 00:00:00
Time taken: 2.03 seconds, Fetched: 10 row(s)
```

7. Creating a Hive-HBase integrated table, denoted as ‘CARD_TRANSACTIONS_HBASE’, ensures accessibility within both Hive and HBase environments.

```
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
```

8. Loading data into ‘CARD_TRANSACTION_HBASE’ table by setting UUID as the primary key for the table.

```
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
```

9. Verifying the table details for “CARD_TRANSACTION_HBASE” and counting the number of rows in base shell.

```
[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>
```

```
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 the relevant data from AWS RDS to Hadoop

1. Importing data into card_member table in RDS to HDFS(ccdf_dir/card_member)

```
[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 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
```

2. Importing data into member_score table in RDS to HDFS(ccdf_dir/member_score)

```
[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
```

3. Creating external table ‘card_member_ext’ to load data from card_member directory in HDFS

```
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
```

4. Creating external table ‘member_score_ext’ to load data from member_score directory in HDFS

```
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
```

5. Creating and loading ‘CARD_MEMBER_ORC' table from the data in ‘card_member_ext table’

```
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. Creating and loading ‘MEMBER_SCORE_ORC’ table from the data in ‘member_score_ext table’

```
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. Retrieving first 10 values from ‘CARD_MEMBER_ORC’ table

```
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)
```

9. Retrieving first 10 values from ‘MEMBER_SCORE_ORC’ table

```
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: Create a look-up table with columns specified in the problem statement

1. Creating an hive-base integrated table named ‘LOOKUP_DATA_HBASE’ from hive shell

```
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
```

2. Verifying the ‘LOOKUP_DATA_HBASE’ attributes

```
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: After creating the table, you need to load the relevant data in the lookup table

1. Creation of ‘RANKED_CARD_TRANSACTION_ORC’ and ‘CARD_UCL_ORC’ table

```
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. Loading data into 'RANKED_CARD_TRANSACTION_ORC'

```
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
```

3. Loading data into ‘CARD_UCL_ORC’ table

```
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
```

4. Loading of the table “LOOKUP_DATA_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
```

5. Validating the record count in “lookup_data_hive” table

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

6. Retrieving 10 records from “LOOKUP_DATA_HBASE” 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
```

7. Validating the data in ‘lookup_data_hive’ in hbase shell

```
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
```