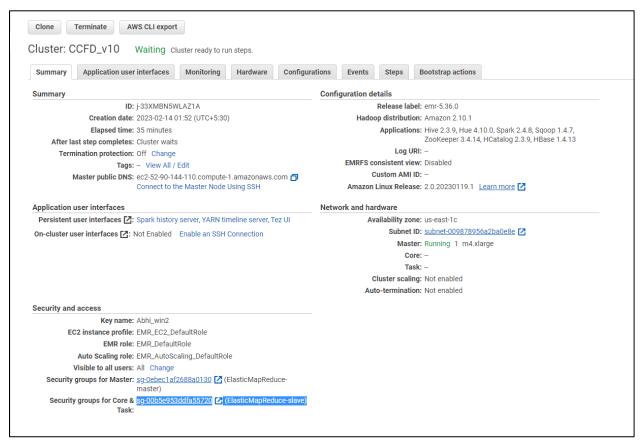# Scripts Execution

Screenshots of the execution of the scripts written

Console Screenshot 1: Created EMR with Hadoop, Hive, HBase, Hcatalog, Spark, Hue and Sqoop.
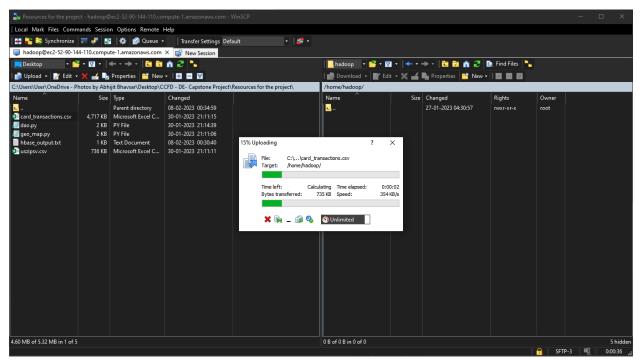
Console Screenshot 2: Created the Database in Hive

Console Screenshot 3: Transferring Historical data to Hadoop local ec2 machine.

Console Screenshot 4: Moving data from EC2 machine to HDFS.

Console Screenshot 5: Creating historical data External table.

```
    > `POSTCODE` STRING,
    > `POS_ID` STRING,
    > `TRANSACTION_DT` STRING,
    > `STATUS` STRING)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    > LOCATION '/user/CCFD_Project/card_transactions.csv'
    > TBLPROPERTIES ("skip.header.line.count"="1");
OK
Time taken: 0.567 seconds
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;
OK
Time taken: 0.875 seconds
hive>
```

Console Screenshot 6: Creating ORC table from external table.

```
Time taken: 0.833 seconds
hive> LOAD DATA INPATH '/user/CCFD_project/card_transactions.csv' INTO TABLE card_transactions_ext;
Loading data to table ccfd.card_transactions_ext
OK
Time taken: 1.066 seconds
hive>
```

Console Screenshot 7: loading data into external table.

```
OK
Time taken: 1.066 seconds
hive> INSERT OVERWRITE TABLE CARD_TRANSACTIONS_ORC
    > SELECT CARD_ID, MEMBER_ID, AMOUNT, POSTCODE, POS_ID, CAST(FROM_UNIXTIME(UNIX_TIMESTAMP(
TION_DT,'dd-MM-yyyy HH:mm:ss')) AS TIMESTAMP), STATUS
    > FROM CARD_TRANSACTIONS_EXT;
Query ID = hadoop_20230213212431_cfce1ea7-f6ce-49b7-85e4-a06e291d55a6
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1676320060972_0003)

--------------------------------------------------------------------------------
        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: 7.49 s
--------------------------------------------------------------------------------
Loading data to table ccfd.card_transactions_orc
OK
Time taken: 12.597 seconds
hive>
```

Console Screenshot 8: Inserting data into orc table.

```
Map 1 .......... container     SUCCEEDED     1       1       0       0       0       0
------------------------------------------------------------------------------------
VERTICES: 01/01  [============================>>] 100%  ELAPSED TIME: 7.49 s
------------------------------------------------------------------------------------
Loading data to table ccfd.card_transactions_orc
OK
Time taken: 12.597 seconds
hive> select count(*) from card_transactions_ext;
Query ID = hadoop_20230213212529_02b24911-f5b5-4152-a3e9-d0b56902b5cc
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1676320060972_0003)

Map 1: 0/1       Reducer 2: 0/1
Map 1: 0(+1)/1   Reducer 2: 0/1
Map 1: 1/1       Reducer 2: 0(+1)/1
Map 1: 1/1       Reducer 2: 1/1
OK
53292
Time taken: 6.029 seconds, Fetched: 1 row(s)
hive>
```

Console Screenshot 9: Validating external table records.


```
Time taken: 0.371 seconds, Fetched: 5 row(s)
hive> SELECT * FROM CARD_TRANSACTIONS_ORC LIMIT 5;
OK
348702330256514 000037495066290 9084849.0     33946   614677375609919 2018-02-11 00:00:00     GENUINE
348702330256514 000037495066290 330148.0      33946   614677375609919 2018-02-11 00:00:00     GENUINE
348702330256514 000037495066290 136052.0      33946   614677375609919 2018-02-11 00:00:00     GENUINE
348702330256514 000037495066290 4310362.0     33946   614677375609919 2018-02-11 00:00:00     GENUINE
348702330256514 000037495066290 9097094.0     33946   614677375609919 2018-02-11 00:00:00     GENUINE
Time taken: 2.017 seconds, Fetched: 5 row(s)
hive>
```

Console Screenshot 10: Validating external table records.

Console Screenshot 11: Validating timestamp column in orc table.

Console Screenshot 12: Creating card_transactions_hbase hive-hbase integrated table.

Console Screenshot 13: Inserting data into from orc table.

Console Screenshot 14: Validating HBase table data.

```
Time taken: 0.192 seconds, Fetched: 10 row(s)
hive> CREATE TABLE LOOKUP_DATA_HBASE(`CARD_ID` STRING,`UCL` DOUBLE, `SCORE` INT, `POSTCODE` STRI
NG, `TRANSACTION_DT` TIMESTAMP)
    > STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
    > WITH SERDEPROPERTIES ("hbase.columns.mapping"=":key, lookup_card_family:ucl, lookup_card_f
amily:score, lookup_transaction_family:postcode, lookup_transaction_family:transaction_dt")
    > TBLPROPERTIES ("hbase.table.name" = "lookup_data_hive");
OK
Time taken: 2.363 seconds
hive>
```
Console Screenshot 15: Creating Lookup Table


```
hive> DESCRIBE lookup_data_hbase;
OK
card_id                 string
ucl                     double
score                   int
postcode                string
transaction_dt          timestamp
Time taken: 0.371 seconds, Fetched: 5 row(s)
hive>
```
Console Screenshot 16:Validating lookup table.


```
hbase(main):001:0> describe 'card_transactions_hive'
Table card_transactions_hive is ENABLED
card_transactions_hive
COLUMN FAMILIES DESCRIPTION
{NAME => 'card_transactions_family', BLOOMFILTER => 'ROW', VERSIONS => '1', IN_MEMORY => 'false', KEEP_DELETED_CELLS => 'F
ALSE', DATA_BLOCK_ENCODING => 'NONE', TTL => 'FOREVER', COMPRESSION => 'NONE', MIN_VERSIONS => '0', BLOCKCACHE => 'true',
BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}
1 row(s) in 0.3570 seconds

hbase(main):002:0>
```
Console Screenshot 17: Checking details of card_transactions_hive hive-HBase integrated table

\

```
hadoop@ip-172-31-26-104:~                                          —   □   ×

hbase(main):002:0> count 'card_transactions_hive'
Current count: 1000, row: 04d77926-52a5-4780-9626-975037fc675f
Current count: 2000, row: 09a98eb7-6765-4388-84c2-188c432d00cd
Current count: 3000, row: 0e9b2a73-c536-4636-a26f-e041eb0a750a
Current count: 4000, row: 1384037c-7476-41a4-8add-97bf6f0f12d3
Current count: 5000, row: 189f4181-834a-41e0-b0be-5a24c2ecf4d4
Current count: 6000, row: 1d7050b2-d6af-423b-961e-b7b9db9973c3
Current count: 7000, row: 2259a721-0b59-412f-b4a6-37b37e334d70
Current count: 8000, row: 272fbc19-125e-494c-85e8-cc7cca5b8942
Current count: 9000, row: 2c5b0d52-c6ac-4eac-a9a0-1d8db30f44cb
Current count: 10000, row: 3158eabb-170b-4fef-b108-35189e3db99a
Current count: 11000, row: 3651744b-90e4-4c3d-8c91-fb536141df03
Current count: 12000, row: 3b07e99c-ee57-4837-971c-745f4082143a
Current count: 13000, row: 3fcb522a-8377-4a66-8dac-10c911c45cc8
Current count: 14000, row: 44b7d9e4-c018-499b-8213-eae818ec072c
Current count: 15000, row: 498d63d8-1c35-486c-851e-d6b031b2606a
Current count: 16000, row: 4e8cee31-ec81-475b-bd0a-36c601597127
Current count: 17000, row: 534486da-adaf-4841-8ce9-0d2151024ecb
Current count: 18000, row: 57d3bf2d-971d-48b6-8c97-4ed51da9fb46
Current count: 19000, row: 5c54041d-089c-4183-9ed5-879e3a09776d
Current count: 20000, row: 60e828db-1c03-4ade-a498-489e2dc66b56
Current count: 21000, row: 659cc46d-0009-4fc9-9891-450a9600ca7b
Current count: 22000, row: 6a565d1b-e0dd-4f21-bb52-73415ee2a612
Current count: 23000, row: 6f15f6dc-49be-4863-9f96-9b44c2d9dbe7
Current count: 24000, row: 7409e355-49db-41eb-9f41-e42a394612f0
Current count: 25000, row: 78907982-d843-4b97-b183-9d2f8afa4053
Current count: 26000, row: 7d8efeb1-d722-4ba9-88e3-88c8673d6571
Current count: 27000, row: 8241c73f-daa2-428b-925d-83c8a47a5dab
Current count: 28000, row: 870d563a-bfd1-42d7-830f-1a50a151a262
Current count: 29000, row: 8bf916fa-652e-4f2d-8b78-c49d22ed847c
Current count: 30000, row: 90c39cca-7ba3-4e28-a1bf-8f710270f385
Current count: 31000, row: 95841e31-e169-4335-bff7-50a474dfdec9
Current count: 32000, row: 9a2f469a-1b08-4f7c-a80e-45c398983344
Current count: 33000, row: 9ee72238-0885-4a1b-868f-8a32a8da0f08
Current count: 34000, row: a36b6d90-0c00-4d38-b8c0-ed68b6f0415d
Current count: 35000, row: a8850ed0-ed9a-4e04-b907-48e03be36677
Current count: 36000, row: ad239cc8-f865-4c8c-b9ea-cf596dbe80f3
Current count: 37000, row: b1f74bea-5ac2-44ba-a821-e6752d25fa1d
Current count: 38000, row: b6a8827d-8151-43d6-86b9-a8bd8412b00f
Current count: 39000, row: bb57a00f-a5a0-4c25-ab05-d40c27dc69cc
Current count: 40000, row: c06b77f9-7b34-40ef-ac76-f558c98a2647
Current count: 41000, row: c5439389-d7da-4f93-b8d9-52893d6f48c4
Current count: 42000, row: c9f3067f-45c3-4e67-8e80-9c8610c823e5
Current count: 43000, row: ceac24b6-f9cb-4795-b338-3328a13a8852
Current count: 44000, row: d3bcb785-f8d6-4404-b68d-8ebc01ddd8a5
Current count: 45000, row: d86f048d-fa9c-4b3e-9470-83571bd21b10
Current count: 46000, row: dd2f3709-79fb-4668-9f8f-e22ebf973499
Current count: 47000, row: e1c87c64-3425-492b-8810-3a394337f6d8
Current count: 48000, row: e6a90bf2-2310-4848-9540-884070b64290
Current count: 49000, row: eb64b7c8-7ac6-4300-88e8-684cafbd48e0
Current count: 50000, row: f0158a52-defa-40a0-af6e-1a2588e8aee8
Current count: 51000, row: f4e850a9-9ccd-4280-90ad-72662756c522
Current count: 52000, row: f9824d77-8879-4af3-b18b-bb89f3859dbd
Current count: 53000, row: fe67799b-54f3-4413-91c7-68c4f14d5c60
53292 row(s) in 4.0110 seconds
```

Console Screenshot 18: Checking the count of the table.

Console Screenshot 19: Checking the details of lookup_data_hive hive-HBase integrated table.

Console Screenshot 20:Altering the lookup_data_hive table and set VERSIONS to 10 for lookup_transaction_family & confirming with describe command.

Console Screenshot 21: For Scoop Import, we will install MySQL connector before starting with Apache Sqoop.

Console Screenshot 22: Scoop import for card member table.

Console Screenshot 23: 999 records imported.

Console Screenshot 24: Scoop import for member score table.

```
            HDFS: Number of write operations=2
        Job Counters
            Launched map tasks=1
            Other local map tasks=1
            Total time spent by all maps in occupied slots (ms)=197328
            Total time spent by all reduces in occupied slots (ms)=0
            Total time spent by all map tasks (ms)=4111
            Total vcore-milliseconds taken by all map tasks=4111
            Total megabyte-milliseconds taken by all map tasks=6314496
        Map-Reduce Framework
            Map input records=999
            Map output records=999
            Input split bytes=87
            Spilled Records=0
            Failed Shuffles=0
            Merged Map outputs=0
            GC time elapsed (ms)=91
            CPU time spent (ms)=1590
            Physical memory (bytes) snapshot=335671296
            Virtual memory (bytes) snapshot=3305467904
            Total committed heap usage (bytes)=326631424
        File Input Format Counters
            Bytes Read=0
        File Output Format Counters
            Bytes Written=19980
23/02/13 21:44:06 INFO mapreduce.ImportJobBase: Transferred 19.5117 KB in 20.2668 seconds (985.8498 bytes/sec)
23/02/13 21:44:06 INFO mapreduce.ImportJobBase: Retrieved 999 records.
[hadoop@ip-172-31-26-104 mysql-connector-java-8.0.25]$
```

Console Screenshot 25: 999 records imported.


```
[hadoop@ip-172-31-26-104 mysql-connector-java-8.0.25]$ hadoop fs -ls /user/CCFD_project/card_member
Found 2 items
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-02-13 21:42 /user/CCFD_project/card_member/_SUCCESS
-rw-r--r--   1 hadoop hdfsadmingroup      85081 2023-02-13 21:42 /user/CCFD_project/card_member/part-m-00000
[hadoop@ip-172-31-26-104 mysql-connector-java-8.0.25]$ hadoop fs -ls /user/CCFD_project/member_score
Found 2 items
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-02-13 21:44 /user/CCFD_project/member_score/_SUCCESS
-rw-r--r--   1 hadoop hdfsadmingroup      19980 2023-02-13 21:44 /user/CCFD_project/member_score/part-m-00000
[hadoop@ip-172-31-26-104 mysql-connector-java-8.0.25]$
```

Console Screenshot 26: Checking the scoop imported files.

Console Screenshot 27: Creating external tables for RDS data.

Console Screenshot 28: Creating ORC tables from the external tables.

Console Screenshot 29: Inserting data into ORC tables from external tables.

```
hadoop@ip-172-31-26-104:~                                    —   □   ✕

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.
properties Async: false
hive> use ccfd;
OK
Time taken: 0.745 seconds
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: 0.855 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: 0.125 seconds
hive>
```

Console Screenshot 30: Creating Ranking ORC table of last 10 customer transactions & Creating Card UCL ORC table for each card.

Console Screenshot 31: Inserting data into ranked_card_transacations_orc table with conditions.

Console Screenshot 32: Inserting data into card_ucl_orc table with conditions.

```
hadoop@ip-172-31-26-104:~                                        —  □  X

Time taken: 8.278 seconds
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;
No Stats for ccfd@ranked_card_transactions_orc, Columns: postcode, rank, transaction_dt, card_id
No Stats for ccfd@card_ucl_orc, Columns: card_id, ucl
No Stats for ccfd@card_member_orc, Columns: member_id, card_id
No Stats for ccfd@member_score_orc, Columns: member_id, score
Query ID = hadoop_20230213223039_af64a4db-ea8c-4511-ab38-4044e72b90df
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1676320060972_0009)

----------------------------------------------------------------------------------------------
        VERTICES        MODE        STATUS    TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container      SUCCEEDED       1         1        0        0       0       0
Map 2 .......... container      SUCCEEDED       1         1        0        0       0       0
Map 3 .......... container      SUCCEEDED       1         1        0        0       0       0
Map 5 .......... container      SUCCEEDED       1         1        0        0       0       0
Reducer 4 ...... container      SUCCEEDED       2         2        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 05/05  [=========================>>] 100%  ELAPSED TIME: 14.04 s
----------------------------------------------------------------------------------------------
OK
Time taken: 21.429 seconds
hive>
```

Console Screenshot 33: Loading data in lookup_data_hbase table.

```
--------------------------------------------------------------------------------
VERTICES: 05/05  [============================>>] 100%  ELAPSED TIME: 14.04 s
--------------------------------------------------------------------------------
OK
Time taken: 21.429 seconds
hive> SELECT COUNT(*) FROM LOOKUP_DATA_HBASE;
Query ID = hadoop_20230213223337_191850c0-0c80-4223-9323-2bc4aa6d5e46
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1676320060972_0009)


--------------------------------------------------------------------------------
        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: 6.39 s
--------------------------------------------------------------------------------
OK
999
Time taken: 9.518 seconds, Fetched: 1 row(s)
hive> SELECT * FROM LOOKUP_DATA_HBASE LIMIT 10;
OK
340028465709212 1.6331555548882348E7     233     24658    2018-01-02 03:25:35
340054675199675 1.4156079786189131E7     631     50140    2018-01-15 19:43:23
340082915339645 1.5285685330791473E7     407     17844    2018-01-26 19:03:47
340134186926007 1.5239767522438556E7     614     67576    2018-01-18 23:12:50
340265728490548 1.608491671255562E7      202     72435    2018-01-21 02:07:35
340268219434811 1.2507323937605347E7     415     62513    2018-01-16 04:30:05
340379737226464 1.4198310998368107E7     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: 0.305 seconds, Fetched: 10 row(s)
hive>
```

Console Screenshot 34: Verifying lookup table data.

```
              at org.jruby.Main.main(Main.java:188)
log4j:ERROR Either File or DatePattern options are not set f
HBase Shell
Use "help" to get list of supported commands.
Use "exit" to quit this interactive shell.
Version 1.4.13, rUnknown, Wed Jun  8 00:30:30 UTC 2022


hbase(main):001:0> count 'lookup_data_hive'
999 row(s) in 0.4720 seconds


=> 999
hbase(main):002:0>
```

Console Screenshot 35: Checking count in lookup_data_hive table.

```
                         079E7
6595928469079750         column=lookup_transaction_family:postcode, timestamp=1676327460635, value=98
                         349
6595928469079750         column=lookup_transaction_family:transaction_dt, timestamp=1676327460635, va
                         lue=2018-01-24 12:38:22
6597703848279563         column=lookup_card_family:score, timestamp=1676327460635, value=218
6597703848279563         column=lookup_card_family:ucl, timestamp=1676327460635, value=1.471863414949
                         8457E7
6597703848279563         column=lookup_transaction_family:postcode, timestamp=1676327460635, value=95
                         699
6597703848279563         column=lookup_transaction_family:transaction_dt, timestamp=1676327460635, va
                         lue=2018-01-27 10:51:49
6598830758632447         column=lookup_card_family:score, timestamp=1676327460635, value=293
6598830758632447         column=lookup_card_family:ucl, timestamp=1676327460635, value=1.222794998260
                         1807E7
6598830758632447         column=lookup_transaction_family:postcode, timestamp=1676327460635, value=19
                         421
6598830758632447         column=lookup_transaction_family:transaction_dt, timestamp=1676327460635, va
                         lue=2018-01-30 00:18:34
6599900931314251         column=lookup_card_family:score, timestamp=1676327460635, value=297
6599900931314251         column=lookup_card_family:ucl, timestamp=1676327460635, value=1.212140857246
                         4656E7
6599900931314251         column=lookup_transaction_family:postcode, timestamp=1676327460635, value=97
                         423
6599900931314251         column=lookup_transaction_family:transaction_dt, timestamp=1676327460635, va
                         lue=2018-01-31 11:25:16
999 row(s) in 2.7220 seconds

hbase(main):003:0>
```

Console Screenshot 36: Checking data in lookup_data_hive table.