

## Loading the Lookup Table

*After creation of the table, loading of the relevant data into the lookup table.*

1. Starting hive session and creating table "ranked\_cc\_transactions\_orc" to store last 10 transactions for eachcard\_id. for better performance.

```
CREATE TABLE IF NOT EXISTS RANKED_CC_TRANSACTIONS_ORC(
`CARD_ID` STRING,
`AMOUNT` DOUBLE,
`POSTCODE` STRING,
`TRANSACTION_DT` TIMESTAMP,
`RANK` INT) STORED AS ORC
TBLPROPERTIES ("orc.compress"="SNAPPY");
```

```
hbase(main):001:0> describe 'cc_lookup_data_hive'
Table cc_lookup_data_hive is ENABLED
cc_lookup_data_hive
COLUMN FAMILIES DESCRIPTION
{NAME => 'lookup_card_family', BLOOMFILTER => 'ROW', VERSIONS => '1', IN_MEMORY => 'false', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', TTL => 'FOREVER', COMPRESSION => 'NONE', MIN_VERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}
{NAME => 'lookup_transactions_family', BLOOMFILTER => 'ROW', VERSIONS => '10', IN_MEMORY => 'false', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', TTL => 'FOREVER', COMPRESSION => 'NONE', MIN_VERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}
2 row(s) in 0.2970 seconds

hbase(main):002:0> exit;
hbase(main):003:0* quit;
hbase(main):004:0* exit
[root@ip-172-31-21-198 ec2-user]# hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> CREATE TABLE IF NOT EXISTS RANKED_CC_TRANSACTIONS_ORC(
> `CARD_ID` STRING,
> `AMOUNT` DOUBLE,
> `POSTCODE` STRING,
> `TRANSACTION_DT` TIMESTAMP,
> `RANK` INT) STORED AS ORC
> TBLPROPERTIES ("orc.compress"="SNAPPY");
OK
Time taken: 1.495 seconds
hive> █
```

2. Create table "cc\_ucl\_orc" to store UCL values for each card\_id. For better performance.

```
CREATE TABLE IF NOT EXISTS CC_UCL_ORC(
`CARD_ID` STRING,
`UCL` DOUBLE) STORED AS ORC
TBLPROPERTIES ("orc.compress"="SNAPPY");
```

```

root@ip-172-31-21-198:/home/ec2-user
bash: describe: command not found
[root@ip-172-31-21-198 ec2-user]# hbase shell
HBase Shell
Use "help" to get list of supported commands.
Use "exit" to quit this interactive shell.
Version 1.4.13, rUnknown, Fri Apr 17 15:18:24 UTC 2020

hbase(main):001:0> describe 'cc_lookup_data_hive'
Table cc_lookup_data_hive is ENABLED
cc_lookup_data_hive
COLUMN FAMILIES DESCRIPTION
{NAME => 'lookup_card_family', BLOOMFILTER => 'ROW', VERSIONS => '1', IN_MEMORY => 'false', KEEP_DELETED
_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', TTL => 'FOREVER', COMPRESSION => 'NONE', MIN_VERSIONS
=> '0', BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}
{NAME => 'lookup_transactions_family', BLOOMFILTER => 'ROW', VERSIONS => '10', IN_MEMORY => 'false', KEE
P_DELETED CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', TTL => 'FOREVER', COMPRESSION => 'NONE', MIN_
VERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}
2 row(s) in 0.2970 seconds

hbase(main):002:0> exit;
hbase(main):003:0* quit;
hbase(main):004:0* exit
[root@ip-172-31-21-198 ec2-user]# hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> CREATE TABLE IF NOT EXISTS RANKED_CC_TRANSACTIONS_ORC(
  > 'CARD_ID' STRING,
  > 'AMOUNT' DOUBLE,
  > 'POSTCODE' STRING,
  > 'TRANSACTION_DT' TIMESTAMP,
  > 'RANK' INT) STORED AS ORC
  > TBLPROPERTIES ("orc.compress"="SNAPPY");
OK
Time taken: 1.495 seconds
hive> CREATE TABLE IF NOT EXISTS CC_UCL_ORC(
  > 'CARD_ID' STRING,
  > 'UCL' DOUBLE) STORED AS ORC
  > TBLPROPERTIES ("orc.compress"="SNAPPY");
OK
Time taken: 0.06 seconds
hive>

```

### 3. Loading data in "ranked\_cc\_transactions\_orc" table

```

INSERT OVERWRITE TABLE RANKED_CC_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, AMOUNT DESC) AS RANK FROM
(SELECT CARD_ID, AMOUNT, POSTCODE, TRANSACTION_DT FROM cc_transactions_HBASE
WHERE STATUS = 'GENUINE') A ) B WHERE B.RANK <= 10;

```

```

root@ip-172-31-21-198:/home/ec2-user
-06-19 17:38:42 GENUINE
0016a375-91c1-4315-b8d0-c77b907573cf 345027679564944 808853092206668 9365301.0 72179 307736895557139 2017-07-20 0
0:27:46 GENUINE
Time taken: 0.244 seconds, Fetched: 10 row(s)
hive> show tables;
OK
card_member_ext
card_transactions_ext
cc_lookup_data_hbase
cc_member_orc
cc_member_score_ext
cc_member_score_orc
cc_transactions_hbase
cc_transactions_orc
cc_ucl_orc
lookup_data_hbase
ranked_cc_transactions_orc
Time taken: 0.02 seconds, Fetched: 11 row(s)
hive> INSERT OVERWRITE TABLE RANKED_CC_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, AMOUNT DESC) AS RANK FROM
> (SELECT CARD_ID, AMOUNT, POSTCODE, TRANSACTION_DT FROM cc_transactions_HBASE
> WHERE STATUS = 'GENUINE') A ) B WHERE B.RANK <= 10;
Query ID = root_20221219191611_8c9d209f-b01b-4326-b59a-b5bd4b010fe7
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1671469134332_0012)

-----
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
-----
VERTICES: 02/02 [=====>>>] 100% ELAPSED TIME: 12.98 s
-----
Loading data to table default.ranked_cc_transactions_orc
OK
Time taken: 17.481 seconds
hive>

```

4. Loading data in "cc\_ucl\_orc" table.  
In the innermost query, select card\_id, average of amount and standard deviation of amount from card\_transactions\_orc.  
In outermost query, select card\_id and compute UCL using average and standard deviation with formula  $(avg + (3 * stddev))$ . Inserting all this data in card\_ucl\_orc.

```

INSERT OVERWRITE TABLE CC_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_CC_TRANSACTIONS_ORC
GROUP BY CARD_ID) A;

```

```

root@ip-172-31-21-198:/home/ec2-user
hive> INSERT OVERWRITE TABLE RANKED_CC_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, AMOUNT DESC) AS RANK FROM
> (SELECT CARD_ID, AMOUNT, POSTCODE, TRANSACTION_DT FROM cc_transactions_HBASE
> WHERE STATUS = 'GENUINE') A ) B WHERE B.RANK <= 10;
Query ID = root_20221219191611_8c9d209f-b01b-4326-b59a-b5bd4b010fe7
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1671469134332_0012)

-----
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
-----
VERTICES: 02/02 [=====>>] 100% ELAPSED TIME: 12.98 s
-----
Loading data to table default.ranked_cc_transactions_orc
OK
Time taken: 17.481 seconds
hive> INSERT OVERWRITE TABLE CC_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_CC_TRANSACTIONS_ORC
> GROUP BY CARD_ID) A;
Query ID = root_20221219192128_ble7ab05-a2f9-450c-8085-1a8237a7ea35
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1671469134332_0012)

-----
VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  RUNNING     1          0          1          0          0          0
Reducer 2 ..... container  INITED      2          0          0          2          0          0
-----
VERTICES: 00/02 [>-----] 0% ELAPSED TIME: 4.07 s
-----

```

5. Load data in **cc\_lookup\_data\_hbase** table.

```

INSERT OVERWRITE TABLE CC_LOOKUP_DATA_HBASE
SELECT RCTO.CARD_ID, CUO.UCL, CMS.SCORE, RCTO.POSTCODE, RCTO.TRANSACTION_DT FROM
RANKED_CC_TRANSACTIONS_ORC RCTO
JOIN CC_UCL_ORC CUO
ON CUO.CARD_ID = RCTO.CARD_ID JOIN (
SELECT DISTINCT CARD.CARD_ID, SCORE.SCORE FROM CC_MEMBER_ORC CARD
JOIN CC_MEMBER_SCORE_ORC SCORE
ON CARD.MEMBER_ID = SCORE.MEMBER_ID) AS CMS ON RCTO.CARD_ID = CMS.CARD_ID
WHERE RCTO.RANK = 1;

```

```

root@ip-172-31-21-198:/home/ec2-user
-----
Map 1 ..... container SUCCEEDED 1 1 0 0 0 0
Reducer 2 ..... container SUCCEEDED 2 2 0 0 0 0
-----
VERTICES: 02/02 [=====]>>] 100% ELAPSED TIME: 6.64 s
-----
Loading data to table default.cc_ucl_orc
OK
Time taken: 7.524 seconds
hive> INSERT OVERWRITE TABLE CC_LOOKUP_DATA_HBASE
> SELECT RCTO.CARD_ID, CUO.UCL, CMS.SCORE, RCTO.POSTCODE, RCTO.TRANSACTION_DT FROM
> RANKED_CC_TRANSACTIONS_ORC RCTO
> JOIN CC_UCL_ORC CUO
> ON CUO.CARD_ID = RCTO.CARD_ID JOIN (
> SELECT DISTINCT CARD.CARD_ID, SCORE.SCORE FROM CC_MEMBER_ORC CARD
> JOIN CC_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 default@ranked_cc_transactions_orc, Columns: postcode, rank, transaction_dt, card_id
No Stats for default@cc_ucl_orc, Columns: card_id, ucl
No Stats for default@cc_member_orc, Columns: member_id, card_id
No Stats for default@cc_member_score_orc, Columns: member_id, score
Query ID = root_20221219192220_7cf9b5e3-100e-4874-b65f-984440d7cc83
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1671469134332_0012)
-----
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: 12.27 s
-----
OK
Time taken: 18.233 seconds
hive>

```

6. Verifying count in “cc\_lookup\_data\_hbase” table.

**select count(\*) from cc\_lookup\_data\_hbase;**

```

root@ip-172-31-21-198:/home/ec2-user
NoViableAltException (240 [])
    at org.apache.hadoop.hive.ql.parse.HiveParser.statement(HiveParser.java:1300)
    at org.apache.hadoop.hive.ql.parse.ParserDriver.parse(ParseDriver.java:208)
    at org.apache.hadoop.hive.ql.parse.ParserUtils.parse(ParseUtils.java:77)
    at org.apache.hadoop.hive.ql.parse.ParserUtils.parse(ParseUtils.java:70)
    at org.apache.hadoop.hive.ql.Driver.compile(Driver.java:468)
    at org.apache.hadoop.hive.ql.Driver.compileInternal(Driver.java:1317)
    at org.apache.hadoop.hive.ql.Driver.runInternal(Driver.java:1457)
    at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1237)
    at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1227)
    at org.apache.hadoop.hive.cli.CliDriver.processLocalCmd(CliDriver.java:233)
    at org.apache.hadoop.hive.cli.CliDriver.processCmd(CliDriver.java:184)
    at org.apache.hadoop.hive.cli.CliDriver.processLine(CliDriver.java:403)
    at org.apache.hadoop.hive.cli.CliDriver.executeDriver(CliDriver.java:821)
    at org.apache.hadoop.hive.cli.CliDriver.run(CliDriver.java:759)
    at org.apache.hadoop.hive.cli.CliDriver.main(CliDriver.java:686)
    at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
    at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
    at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
    at java.lang.reflect.Method.invoke(Method.java:498)
    at org.apache.hadoop.util.RunJar.run(RunJar.java:239)
    at org.apache.hadoop.util.RunJar.main(RunJar.java:153)
FAILED: ParseException line 1:8 cannot recognize input near 'VERTICES' 'exit' '<EOF>'
hive> select count(*) from cc_lookup_data_hbase;
Query ID = root_20221219192420_311bfaf17-616d-4a42-b4ae-159e9d9df720f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1671469134332_0012)
-----
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: 6.60 s
-----
OK
999
Time taken: 11.158 seconds, Fetched: 1 row(s)
hive>

```



Total number of records is **999** which is matching with given requirement.

- Verify some data in “cc\_lookup\_data\_hbase” table.

**select \* from cc\_lookup\_data\_hbase limit 10;**

```
root@ip-172-31-21-198:/home/ec2-user
at org.apache.hadoop.hive.cli.CliDriver.executeDriver(CliDriver.java:821)
at org.apache.hadoop.hive.cli.CliDriver.run(CliDriver.java:759)
at org.apache.hadoop.hive.cli.CliDriver.main(CliDriver.java:686)
at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:498)
at org.apache.hadoop.util.RunJar.run(RunJar.java:239)
at org.apache.hadoop.util.RunJar.main(RunJar.java:153)
FAILED: ParseException line 1:8 cannot recognize input near 'VERTICES' 'exit' '<EOF>'
hive> select count(*) from cc_lookup_data_hbase;
Query ID = root_20221219192420_311bfa17-616d-4a42-b4ae-159e9ddf720f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1671469134332_0012)

-----
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: 6.60 s
-----
OK
999
Time taken: 11.158 seconds, Fetched: 1 row(s)
hive> select * from cc_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: 0.192 seconds, Fetched: 10 row(s)
hive>
```

- Start HBase shell session and verifying count in “cc\_lookup\_data\_hive” table.

**count 'cc\_lookup\_data\_hive'**

```
hbase(main):002:0> count 'cc_lookup_data_hive'
999 row(s) in 0.3560 seconds
=> 999
hbase(main):003:0>
```

Total number of records is **999** which is matching with given requirement.

- Verifying data in “cc\_lookup\_data\_hive” table.

**scan 'cc\_lookup\_data\_hive'**

```

root@ip-172-31-21-198:/home/ec2-user
6592184145413632 column=lookup_transactions_family:postcode, timestamp=1671477758927, value=53186
6592184145413632 column=lookup_transactions_family:transactions_dt, timestamp=1671477758927, value=2018-01-2
8 00:54:30
6594248319343442 column=lookup_card_family:score, timestamp=1671477758927, value=350
6594248319343442 column=lookup_card_family:ucl, timestamp=1671477758927, value=1.4567957140418548E7
6594248319343442 column=lookup_transactions_family:postcode, timestamp=1671477758927, value=24927
6594248319343442 column=lookup_transactions_family:transactions_dt, timestamp=1671477758927, value=2018-01-3
1 23:42:38
6595638658736751 column=lookup_card_family:score, timestamp=1671477758927, value=310
6595638658736751 column=lookup_card_family:ucl, timestamp=1671477758927, value=1.356629177577566E7
6595638658736751 column=lookup_transactions_family:postcode, timestamp=1671477758927, value=68328
6595638658736751 column=lookup_transactions_family:transactions_dt, timestamp=1671477758927, value=2018-01-3
0 10:50:34
6595814135833988 column=lookup_card_family:score, timestamp=1671477758927, value=210
6595814135833988 column=lookup_card_family:ucl, timestamp=1671477758927, value=1.3926273240525039E7
6595814135833988 column=lookup_transactions_family:postcode, timestamp=1671477758927, value=22508
6595814135833988 column=lookup_transactions_family:transactions_dt, timestamp=1671477758927, value=2018-01-3
0 02:03:54
6595928469079750 column=lookup_card_family:score, timestamp=1671477758927, value=412
6595928469079750 column=lookup_card_family:ucl, timestamp=1671477758927, value=1.142797041440079E7
6595928469079750 column=lookup_transactions_family:postcode, timestamp=1671477758927, value=98349
6595928469079750 column=lookup_transactions_family:transactions_dt, timestamp=1671477758927, value=2018-01-2
4 12:38:22
6597703848279563 column=lookup_card_family:score, timestamp=1671477758927, value=218
6597703848279563 column=lookup_card_family:ucl, timestamp=1671477758927, value=1.4718634149498457E7
6597703848279563 column=lookup_transactions_family:postcode, timestamp=1671477758927, value=95699
6597703848279563 column=lookup_transactions_family:transactions_dt, timestamp=1671477758927, value=2018-01-2
7 10:51:49
6598830758632447 column=lookup_card_family:score, timestamp=1671477758927, value=293
6598830758632447 column=lookup_card_family:ucl, timestamp=1671477758927, value=1.2227949982601807E7
6598830758632447 column=lookup_transactions_family:postcode, timestamp=1671477758927, value=19421
6598830758632447 column=lookup_transactions_family:transactions_dt, timestamp=1671477758927, value=2018-01-3
0 00:18:34
6599900931314251 column=lookup_card_family:score, timestamp=1671477758927, value=297
6599900931314251 column=lookup_card_family:ucl, timestamp=1671477758927, value=1.2121408572464656E7
6599900931314251 column=lookup_transactions_family:postcode, timestamp=1671477758927, value=97423
6599900931314251 column=lookup_transactions_family:transactions_dt, timestamp=1671477758927, value=2018-01-3
1 11:25:16
999 row(s) in 5.6360 seconds
hbase(main):006:0>

```