Loading Historical Transactions Data into NoSQL Database

Commands to load the past transactions data into NoSQL database

1. Creating External table in 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
'/user/CCFD_Project/card_transactions.csv' TBLPROPERTIES
("skip.header.line.count"="1");
```

2. Loading the data from to external hive table

```
LOAD DATA LOCAL INPATH 'CARD_TRANSACTIONS.CSV' INTO TABLE
CARD_TRANSACTIONS_EXT;
```

3. Converting external table to ORC provides a more efficient and optimized data storage and retrieval solution for 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");
```

4. Inserting data into ORC table

```
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;
```

5. Creating hive-hbase integrated table which will be visible in HBase as well

```
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");
```

6. Inserting data into Hive-Hbase integrated table

```
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;
```

Command to list the table in which the data is loaded and the command to get the count of the rows of the table

1. Verifying the data for CARD_TRANSACTIONS_EXT

```
SELECT COUNT(*) FROM CARD_TRANSACTIONS_EXT;
```

2. Verifying the data for CARD_TRANSACTIONS_HBASE

```
SELECT * FROM CARD_TRANSACTIONS_HBASE LIMIT 10;
```

3. Verifying the data for CARD_TRANSACTIONS_HBASE

```
SELECT * FROM CARD_TRANSACTIONS_HBASE LIMIT 10;
```

4. Verifying if the timestamp is working properly

```
SELECT YEAR(TRANSACTION_DT),TRANSACTION_DT FROM
CARD_TRANSACTIONS_ORC LIMIT 10;
```

Screenshot of the table created

```
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 '/user/CCFD_Project/card_transactions.csv'
    > TBLPROPERTIES ("skip.header.line.count"="1");
OK
Time taken: 0.348 seconds
hive>
```

Console Screenshot 1:Creating historical data External table.

```
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: 0.687 seconds
```

Console Screenshot 2: loading data into external table

```
hive> INSERT OVERWRITE TABLE CARD_TRANSACTIONS_ORC
    > SELECT CARD_ID, MEMBER_ID, AMOUNT, POSTCODE, POS_ID, CAST(FROM_UNIXTIME(UNIX_TIMESTAMP(TRANSACTION_DT,'dd-M
M-yyyy HH:mm:ss')) AS TIMESTAMP), STATUS FROM CARD_TRANSACTIONS_EXT;
Query ID = hadoop_20250920100844_0879aa6e-baec-4687-985e-6ab3eb01d361
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1758361848317_0001)

----------------------------------------------------------------------------------------------
        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: 7.82 s
----------------------------------------------------------------------------------------------
Loading data to table ccfd.card_transactions_orc
OK
Time taken: 11.342 seconds
hive>
```

Console Screenshot : overwriting data into external 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;
OK
Time taken: 0.516 seconds
hive> show tables;
OK
card_transactions_ext
card_transactions_orc
Time taken: 0.049 seconds, Fetched: 2 row(s)
hive>
```

Console Screenshot 3: Creating ORC table from external table

```
hive> INSERT OVERWRITE TABLE CARD_TRANSACTIONS_ORC
    > SELECT CARD_ID, MEMBER_ID, AMOUNT, POSTCODE, POS_ID, CAST(FROM_UNIXTIME(UNIX_TIMESTAMP(TRANSACTION_DT,'dd-M
M-yyyy HH:mm:ss')) AS TIMESTAMP), STATUS FROM CARD_TRANSACTIONS_EXT;
Query ID = hadoop_20250920100844_0879aa6e-baec-4687-985e-6ab3eb01d361
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1758361848317_0001)

--------------------------------------------------------------------------------
        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: 7.82 s
--------------------------------------------------------------------------------
Loading data to table ccfd.card_transactions_orc
OK
Time taken: 11.342 seconds
hive>
```

Console Screenshot 4: Inserting data into orc table

```
hive> select count(*) from card_transactions_ext;
Query ID = hadoop_20250920101007_29b3c0ee-b554-44df-a69c-6a8f757b8212
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1758361848317_0001)

--------------------------------------------------------------------------------
        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: 3.86 s
--------------------------------------------------------------------------------
OK
53292
Time taken: 6.205 seconds, Fetched: 1 row(s)
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: 0.125 seconds, Fetched: 10 row(s)
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: 3.226 seconds
hive>
```

Console Screenshot 5: Creating card_transactions_hbase hive-hbase integrated 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_20250920101143_d064a84d-024a-4cc1-a5cb-940645fc81b5
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1758361848317_0001)

--------------------------------------------------------------------------------
        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: 8.47 s
--------------------------------------------------------------------------------
OK
Time taken: 13.923 seconds
hive>
```

Console Screenshot 6: Inserting data into from orc table.

```
hive> select * from card_transactions_hbase limit 10;
OK
000008b9-7ba0-490e-a858-215a44c9534a    5391723993945313        997128952368160 928092.0        15139   966249108
897809  2017-12-26 23:57:23     GENUINE
00006283-b90b-490e-9767-016b1f5b5049    5380072688020054        036571958569825 7789248.0       78349   750827915
152480  2018-01-22 18:30:19     GENUINE
00014053-9526-4a4b-91b5-a9bc4542d097    344002520206946 313129704156102 2326643.0       29456   418578505560437 2
016-08-15 23:50:35      GENUINE
0001a103-c310-491b-92f6-f4f06a8fa4ef    4540807128933493        241809163782996 4902721.0       76649   170198462
440301  2017-04-17 07:32:28     GENUINE
00020a46-e50c-4950-962d-9d2c16d565c5    349143706735646 343824445342591 4813100.0       89439   274727493822152 2
017-07-20 18:01:08      GENUINE
0002fe73-ab94-4fff-88fc-f9df935a7180    4314008605559737        846409651699691 845437.0        45880   465558985
506070  2017-03-02 23:04:14     GENUINE
000614e8-c98a-4b09-b22f-7374e2930856    6011544071439690        583773644956274 2846891.0       58544   801035095
940538  2017-12-31 23:20:13     GENUINE
00072f0d-7a73-4509-8426-e3641c134f93    6011740360743178        066377656985754 1803531.0       65032   779289002
769653  2017-07-16 08:40:58     GENUINE
0008609c-5062-4abe-b24f-c647c0db80f3    348890647161465 320864375768815 589728.0        49797   710253657063163 2
018-01-16 09:46:05      GENUINE
000960d1-2a48-4066-b8b6-08d64e59fbdf    6228588350544786        669696032320589 3181868.0       19512   448946531
936650  2016-10-07 07:05:44     GENUINE
Time taken: 0.147 seconds, Fetched: 10 row(s)
hive>
```