Loading Historical Transactions Data into NoSQL Database

Commands to load the past transactions data into NoSQL database

---------- Hive Operations: Starts Here ----------

1. Start hive and create new database named **ccfd_capstone_project** -> switch to **ccfd_capstone_project** database

   ```
   create database ccfd_capstone_project;
   use ccfd_capstone_project;
   ```

2. Set below parameters for the hive session

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

3. Create an external table "**card_transactions_ext**" and store it in /ccfd_capstone_project/card_transactions

   ```
   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_capstone_project/card_transactions'
   TBLPROPERTIES ("skip.header.line.count"="1");
   ```

4. Create table "**card_transactions_orc**" in ORC format for better performance.

```
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");
```

5. Create hive-hbase integrated table which will be visible in HBase as well. "**card_transactions_hbase**" table

```
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");
```

---------- Hive Operations: Ends Here ----------

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

---------- Hive Operations: Starts Here ----------

1. Load data in “**card_transactions_orc**” table and type cast transaction_dt column in timestamp format

   ```
   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;
   ```

2. Verify transaction_dt and year columns in "**card_transactions_orc**" table.

   ```
   select year(transaction_dt), transaction_dt from card_transactions_orc limit 10;
   ```

3. Load data in "**card_transactions_hbase**" table which will be visible in HBase as well with table name as "**card_transactions_hive**".Using randomUUID to populate TRANSACTION_ID field (row key).

   ```
   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;
   ```

4. Verify data in "**card_transactions_hbase**" table.

   select * from card_transactions_hbase limit 10;

---------- Hive Operations: Ends Here ----------

---------- Hbase Operations: Starts Here ----------

1. Start HBase and verify details of "**card_transactions_hive**" table (hive-hbase integrated table).

   describe 'card_transactions_hive'

2. Verify count of "**card_transactions_hive**" table Command :

   count 'card_transactions_hive'

---------- Hbase Operations: Ends Here ----------

Screenshot of the table created

1. create new database named ccfd_capstone_project and set parameters

```
hive> create database ccfd_capstone_project;
OK
Time taken: 0.808 seconds
hive>
```

```
hive> use ccfd_capstone_project;
OK
Time taken: 0.029 seconds
hive>
```

```
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;
hive> 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;
hive>
```

2. Create an external table "**card_transactions_ext**"

```
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_capstone_project/card_transactions' TBLPROPERTIES ("skip.header.line.count"="1");
OK
Time taken: 0.595 seconds
hive>
```

3. Create table "**card_transactions_orc**"

```
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: 0.339 seconds
hive>
```

4. Load data in “**card_transactions_orc**”

```
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 = root_20230101213340_a1ef810f-085a-4f7a-a62c-009278f9c6c5
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1672606124138_0004)

--------------------------------------------------------------------------------
        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: 5.63 s
--------------------------------------------------------------------------------
Loading data to table default.card_transactions_orc
OK
Time taken: 9.89 seconds
hive>
```

5. Verify **transaction_dt** and year columns in "**card_transactions_orc**" table.

```
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: 0.326 seconds, Fetched: 10 row(s)
hive>
```

6. Create hive-hbase integrated table which will be visible in HBase as well. "**card_transactions_hbase**" table

```
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: 2.637 seconds
hive>
```

7. Load data in "**card_transactions_hbase**".

```
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 = root_20230101213746_7ac492a7-cbf8-432e-aacc-392668a12eb7
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1672606124138_0004)

----------------------------------------------------------------------------------------------
        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: 7.95 s
----------------------------------------------------------------------------------------------
OK
Time taken: 10.145 seconds
hive>
```

8. Verify data in "card_transactions_hbase" table.

```
hive> select * from card_transactions_hbase limit 10;
OK
0000702c-cf3a-4cc5-8bd9-5bc34c32b674    6451188046445957        482846848859991 4765412.0       17814   980020874687881 2017-04-02 18:10:37     GENUINE
0000b0ae-237d-4280-a101-88f5fd857877    5556186648549560        798053888675530 5346125.0       22547   805601786810909 2017-11-11 00:00:00     GENUINE
00017f59-7f0e-40d0-9069-bc4ee7c40e23    5360062424232117        895011420197476 1979385.0       25866   415503630323997 2017-09-23 22:15:24     GENUINE
0007913a-7d87-4b55-9bd5-ae667b987353    6011082928436197        069021032902949 5674514.0       49738   217648815673353 2016-11-06 06:53:28     GENUINE
0009622b-8b6f-481b-aedd-3a77f484f498    375773536539674 146862049588235 3953206.0       29567   535569409136312 2018-01-03 07:30:57     GENUINE
000aec26-549f-4865-971b-4446c11b3536    6011938409004772        577907767500023 3286335.0       18943   555922206644053 2018-01-10 22:05:54     GENUINE
000b7048-bdd8-4785-b2d2-99208f3e5450    375372047396189 595995388849040 8089858.0       98243   604211641417664 2017-12-01 02:28:28     GENUINE
000d0d23-84d7-4d94-938e-c0194f5c9336    5589613730225354        054411454572492 6680194.0       12033   605815588589423 2018-01-31 00:53:16     GENUINE
000d483c-85fe-4712-86fb-4c9be734d555    6440187483823803        056816206595507 244334.0        98020   641700902956399 2017-03-13 15:23:16     GENUINE
000e3a31-dbef-48be-b484-979513b82f95    5127318999406559        391603008295007 1282764.0       26058   357112280203781 2017-08-17 04:22:05     GENUINE
Time taken: 0.291 seconds, Fetched: 10 row(s)
hive>
```

9. verify details of "**card_transactions_hive**" 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', M
IN_VERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}

1 row(s) in 0.3010 seconds

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

10. Verify count of "**card_transactions_hive**"

```
Current count: 30000, row: 90baa19d-a7c9-4309-bf3b-74be41f4375a
Current count: 31000, row: 956b49ea-6605-4fba-a9ed-3e1dc34ed7a7
Current count: 32000, row: 9a2b7792-3a77-4da7-b1d8-f2572de10abf
Current count: 33000, row: 9edc7672-ab66-4c53-a956-c80cfe09c970
Current count: 34000, row: a400fc06-3a54-4c6a-a0da-5ad0ca8e3683
Current count: 35000, row: a8d95a8e-e4bb-4a73-9740-15db43858ad4
Current count: 36000, row: ada63fa5-365e-4bf6-acfd-10bff82a52c7
Current count: 37000, row: b264efa7-a23f-4a0d-8dcb-0ab715f5261a
Current count: 38000, row: b6fd7e53-d3fb-4f6c-9443-91130c58b68d
Current count: 39000, row: bbb93af5-4c48-4ac0-b384-b480be67174b
Current count: 40000, row: c0751f1d-3620-434c-b72a-c56ae8bdbd26
Current count: 41000, row: c4db5d8b-63a7-40c8-9f80-743fc5e5ce00
Current count: 42000, row: c9b43e40-b5a2-4660-92ae-535ef1601d39
Current count: 43000, row: ce67a599-2d57-4b37-8fe0-b3e5e19bf2f7
Current count: 44000, row: d2fc23eb-9ccb-4ed9-859c-cf99c3bf66a9
Current count: 45000, row: d7a39468-e864-4409-8850-58db409437c3
Current count: 46000, row: dc878ff7-f740-490c-b693-e6543e089e6a
Current count: 47000, row: e167adc9-9d02-42e9-b0ae-658cdb28ec4b
Current count: 48000, row: e6475751-9e59-4134-87bd-f2cde1a2bef8
Current count: 49000, row: eb1e895f-f278-4ff2-978b-1416a0a3686e
Current count: 50000, row: f0633581-76fe-4425-bc1c-dceeba4d3423
Current count: 51000, row: f509c828-3a2c-48d7-a5b5-f315b5aeae59
Current count: 52000, row: f9d2dd4d-09d0-4b4e-9717-8ec3556b74d5
Current count: 53000, row: fe8de1bd-7d0f-4fb6-a05a-3553eb879f93
53292 row(s) in 2.8390 seconds

=> 53292
hbase(main):004:0>
```

Count of the "**card_transactions_hive**" table is **53292** which is matching with given requirement.