Loading Historical Transactions Data into NoSQL Database

Commands to load the past transactions data into NoSQL database

- Push provided csv file to HDFS
  - *hadoop fs -mkdir /user/capstone*
  - *hadoop fs -put card_transactions.csv /user/capstone/card_transaction*

- Create database and table in hive
  - *create database capstone;*
  - *use capstone;*
  - *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/capstone/card_transaction'
TBLPROPERTIES ("skip.header.line.count"="1");
```

  - Check count of rows loaded

    SELECT COUNT(*) FROM card_transactions_ext;

- We observed that records were not sorting when doing sort on transaction date column, hence moved the data to another table with formatting.
  - Create intermediate table to store formatted data

```
CREATE TABLE IF NOT EXISTS transactions_formatted (
    `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");

- Format and load into intermediate table

INSERT OVERWRITE TABLE transactions_formatted
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;

Screenshot of the table created

- Pushing files to HDFS

```
[hadoop@ip-172-31-77-115 ~]$ hadoop fs -put capstone_project/FileTransactions/card_transactions.csv /user/capstone/card_transaction
[hadoop@ip-172-31-77-115 ~]$ hadoop fs -ls /user/capstone/card_transaction
Found 1 items
-rw-r--r--   1 hadoop hadoop    4829520 2023-12-09 06:36 /user/capstone/card_transaction/card_transactions.csv
[hadoop@ip-172-31-77-115 ~]$ hive
```

- Screenshot of table creation and count of loaded rows

```
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
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/capstone/card_transaction'
    > TBLPROPERTIES ("skip.header.line.count"="1");
OK
Time taken: 0.695 seconds
hive> select count(*) from card_transactions_ext;
Query ID = hadoop_20231209063807_396e46f2-8da6-4e1a-bc2c-2c1563ca0315
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1702099609810_0011)

--------------------------------------------------------------------------------
        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: 5.22 s
--------------------------------------------------------------------------------
OK
53292
Time taken: 8.774 seconds, Fetched: 1 row(s)
```

Expected Result -> Loaded 53292 rows successfully

- Create intermediate table to store formatted data

```
hive> CREATE TABLE IF NOT EXISTS transactions_formatted (
    > `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: 0.06 seconds
hive>
```

- Format and copy data from main table to intermediate one for later use

```
Time taken: 0.06 seconds
hive> INSERT OVERWRITE TABLE transactions_formatted
    > 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_20231209095602_edc649b9-5ad8-4952-8b11-f59af280840f
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1702114013497_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: 7.79 s
--------------------------------------------------------------------------------
Loading data to table capstone.transactions_formatted
OK
Time taken: 16.095 seconds
hive>
```