Scripts Execution

Following is the sequence of steps executed in order.

Task 1: Load the transactions history data (card_transactions.csv) in a NoSQL database

1. Create and upload the Transactions file first to hadoop cluster in order to upload to NoSQL database.
   Command to do execute:
   hadoop fs -mkdir /apps/capstone
   hadoop fs -mkdir /apps/capstone/card_transaction
   hadoop fs -put card_transactions.csv /apps/capstone/card_transaction/

```
/card_transaction/
[hadoop@ip-172-31-69-191 ~]$ hadoop fs -ls /apps/capstone/card_transaction/
Found 1 items
-rw-r--r--   1 hadoop hadoop    4829520 2023-12-24 05:47 /apps/capstone/card_tra
nsaction/card_transactions.csv
[hadoop@ip-172-31-69-191 ~]$
```

2. We have selected the hive database so on hadoop cluster connect to hive.
   a. User command 'hive'
      Create database capstone;
      Use capstone;

- **Task 2**: Ingest the relevant data from AWS RDS to Hadoop.

3. Create table using command:

```
Time taken: 0.115 seconds
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 '/apps/capstone/card_transaction'
    > TBLPROPERTIES ("skip.header.line.count"="1");
OK
Time taken: 0.367 seconds
hive> select count(*) from card_transactions_ext;
Query ID = hadoop_20231224055007_e564b8e7-205c-4370-b2d3-d88ee1cb43aa
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1703396126878
_0001)

Map 1: 0/1      Reducer 2: 0/1
Map 1: 0/1      Reducer 2: 0/1
Map 1: 0(+1)/1  Reducer 2: 0/1
Map 1: 0/1      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: 11.278 seconds, Fetched: 1 row(s)
hive>
```

4. Create another table to handle date time datatyping issue and move from external table to internal table.

```
Time taken: 11.278 seconds, Fetched: 1 row(s)
hive> CREATE TABLE IF NOT EXISTS transactions_formatted (
    > TRANSACTION_ID STRING,
    > CARD_ID STRING,
    > MEMBER_ID STRING,
    > AMOUNT DOUBLE,
    > POSTCODE STRING,
    > POS_ID STRING,
    > TRANSACTION_DT TIMESTAMP,
    > STATUS STRING)
    > STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
    > WITH SERDEPROPERTIES ("hbase.columns.mapping"=":key, transactions:CARD_ID, transactions:MEMBER_ID, transactions:AMOUNT,transactions:POSTCODE, transactions:POS_ID, transactions:TRANS
TION_DT, transactions:STATUS")
    > TBLPROPERTIES ("hbase.table.name" = "transaction_hbase");
OK
Time taken: 2.834 seconds
hive>
```

5. Move data using script:

```
Time taken: 2.834 seconds
hive> INSERT OVERWRITE TABLE transactions_formatted
    > SELECT reflect('java.util.UUID', 'randomUUID') as TRANSACTION_ID,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_20231224055157_da073fad-8d80-468b-8033-425fdc746369
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1703396126878_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: 11.01 s
--------------------------------------------------------------------------------
OK
Time taken: 14.86 seconds
hive>
```

6. Next step is to do Sqoop job to import Member score and card member data from AMS RDS job to required location.

    a. Create the folders for these jobs using

i. hadoop fs -mkdir /user/capstone/card_member
ii. hadoop fs -mkdir /user/capstone/member_score

b. Set hive parameter and snappy config paramters for fast performance:

Set Hive parameters:

```
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;
```

c. Sqoop Job Commands:

```
sqoop import --connect jdbc:mysql://upgradawsrds1.cyaielc9bmnf.us-east-1.rds.amazonaws.com:3306/cred_financials_data --username upgraduser --password upgraduser --table member_score --null-string 'NA' --null-non-string '\\N' --delete-target-dir --target-dir '/apps/capstone/member_score' -m 1
```

```
[hadoop@ip-172-31-77-115 ~]$ sqoop import --connect jdbc:mysql://upgradawsrds1.cyaielc9bmnf.us-east-1.rds.amazonaws.com:3306/cred_financials_data --username upgraduser --password upgraduser
 --table member_score --null-string 'NA' --null-non-string '\\N' --delete-target-dir --target-dir '/user/capstone/member_score' -m 1
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
23/12/09 06:58:46 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/aws/redshift/jdbc/redshift-jdbc42-1.2.37.1061.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hive/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
23/12/09 06:58:46 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
23/12/09 06:58:46 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
23/12/09 06:58:46 INFO tool.CodeGenTool: Beginning code generation
23/12/09 06:58:47 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `member_score` AS t LIMIT 1
23/12/09 06:58:47 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `member_score` AS t LIMIT 1
23/12/09 06:58:47 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-hadoop/compile/2aa950db0035bc0b1599cec045e0f7e2/member_score.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
23/12/09 06:58:50 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hadoop/compile/2aa950db0035bc0b1599cec045e0f7e2/member_score.jar
23/12/09 06:58:51 INFO tool.ImportTool: Destination directory /user/capstone/member_score is not present, hence not deleting.
23/12/09 06:58:51 WARN manager.MySQLManager: It looks like you are importing from mysql.
23/12/09 06:58:51 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
23/12/09 06:58:51 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
23/12/09 06:58:51 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
23/12/09 06:58:51 INFO mapreduce.ImportJobBase: Beginning import of member_score
23/12/09 06:58:51 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
23/12/09 06:58:51 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
23/12/09 06:58:51 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-77-115.ec2.internal/172.31.77.115:8032
23/12/09 06:58:52 WARN hdfs.DataStreamer: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1257)
        at java.lang.Thread.join(Thread.java:1331)
        at org.apache.hadoop.hdfs.DataStreamer.closeResponder(DataStreamer.java:973)
        at org.apache.hadoop.hdfs.DataStreamer.endBlock(DataStreamer.java:624)
        at org.apache.hadoop.hdfs.DataStreamer.run(DataStreamer.java:801)
23/12/09 06:58:53 INFO db.DBInputFormat: Using read commited transaction isolation
23/12/09 06:58:53 INFO mapreduce.JobSubmitter: number of splits:1
23/12/09 06:58:53 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1702099609810_0013
23/12/09 06:58:53 INFO impl.YarnClientImpl: Submitted application application_1702099609810_0013
23/12/09 06:58:53 INFO mapreduce.Job: The url to track the job: http://ip-172-31-77-115.ec2.internal:20888/proxy/application_1702099609810_0013/
23/12/09 06:58:53 INFO mapreduce.Job: Running job: job_1702099609810_0013
23/12/09 06:59:01 INFO mapreduce.Job: Job job_1702099609810_0013 running in uber mode : false
23/12/09 06:59:01 INFO mapreduce.Job:  map 0% reduce 0%
```

```
23/12/09 06:59:01 INFO mapreduce.Job: Job job_1702099609810_0013 running in uber mode : false
23/12/09 06:59:01 INFO mapreduce.Job:  map 0% reduce 0%
23/12/09 06:59:07 INFO mapreduce.Job:  map 100% reduce 0%
23/12/09 06:59:08 INFO mapreduce.Job: Job job_1702099609810_0013 completed successfully
23/12/09 06:59:08 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=189991
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=87
                HDFS: Number of bytes written=19980
                HDFS: Number of read operations=4
                HDFS: Number of large read operations=0
                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)=173376
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=3612
                Total vcore-milliseconds taken by all map tasks=3612
                Total megabyte-milliseconds taken by all map tasks=5548032
        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)=78
                CPU time spent (ms)=1750
                Physical memory (bytes) snapshot=260579328
                Virtual memory (bytes) snapshot=3286204416
                Total committed heap usage (bytes)=246939648
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=19980
23/12/09 06:59:08 INFO mapreduce.ImportJobBase: Transferred 19.5117 KB in 17.2766 seconds (1.1294 KB/sec)
23/12/09 06:59:08 INFO mapreduce.ImportJobBase: Retrieved 999 records.
[hadoop@ip-172-31-77-115 ~]$
```

sqoop import --connect jdbc:mysql://upgradawsrds1.cyaielc9bmnf.us-east-1.rds.amazonaws.com:3306/cred_financials_data --username upgraduser --password upgraduser --table card_member --null-string 'NA' --null-non-string '\\N' --delete-target-dir --target-dir '/apps/capstone/card_member' -m 1

```
hadoop@ip-172-31-77-115:~
23/12/09 07:01:41 INFO db.DBInputFormat: Using read commited transaction isolation
23/12/09 07:01:41 INFO mapreduce.JobSubmitter: number of splits:1
23/12/09 07:01:41 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1702099609810_0014
23/12/09 07:01:42 INFO impl.YarnClientImpl: Submitted application application_1702099609810_0014
23/12/09 07:01:42 INFO mapreduce.Job: The url to track the job: http://ip-172-31-77-115.ec2.internal:20888/proxy/application_1702099609810_0014/
23/12/09 07:01:42 INFO mapreduce.Job: Running job: job_1702099609810_0014
23/12/09 07:01:51 INFO mapreduce.Job: Job job_1702099609810_0014 running in uber mode : false
23/12/09 07:01:51 INFO mapreduce.Job:  map 0% reduce 0%
23/12/09 07:01:58 INFO mapreduce.Job:  map 100% reduce 0%
23/12/09 07:01:58 INFO mapreduce.Job: Job job_1702099609810_0014 completed successfully
23/12/09 07:01:58 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=190043
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=87
                HDFS: Number of bytes written=85081
                HDFS: Number of read operations=4
                HDFS: Number of large read operations=0
                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)=224976
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=4687
                Total vcore-milliseconds taken by all map tasks=4687
                Total megabyte-milliseconds taken by all map tasks=7199232
        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)=87
                CPU time spent (ms)=2740
                Physical memory (bytes) snapshot=279830528
                Virtual memory (bytes) snapshot=3280777216
                Total committed heap usage (bytes)=249036800
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=85081
23/12/09 07:01:58 INFO mapreduce.ImportJobBase: Transferred 83.0869 KB in 18.4139 seconds (4.5122 KB/sec)
23/12/09 07:01:58 INFO mapreduce.ImportJobBase: Retrieved 999 records.
[hadoop@ip-172-31-77-115 ~]$
```

7. Create hive tables from the loaded tables from the sqoop data Ingestions results:

Member_score table:

```
[hadoop@ip-172-31-69-191 ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> use capstone;
OK
Time taken: 0.691 seconds
hive> CREATE EXTERNAL TABLE IF NOT EXISTS member_score(
    > MEMBER_ID String,
    > score String)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    > LINES TERMINATED BY '\n'
    > LOCATION '/apps/capstone/member_score';
OK
Time taken: 0.327 seconds
hive> select count(*) from member_score;
Query ID = hadoop_20231224055728_cf129a57-5b6b-42bb-ac1a-11ea884b11f3
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1703396126878_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reduce 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 4.84 s
--------------------------------------------------------------------------------
OK
999
Time taken: 8.922 seconds, Fetched: 1 row(s)
hive>
```

Card_member table

```
Time taken: 8.922 seconds, Fetched: 1 row(s)
hive> CREATE EXTERNAL TABLE IF NOT EXISTS card_member(
    > card_id string,
    > MEMBER_ID string,
    > score string,
    > tr_date string,
    > exp_date string,
    > country string,
    > area string)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    > LINES TERMINATED BY '\n'
    > LOCATION '/apps/capstone/card_member';
OK
Time taken: 0.106 seconds
hive> select count(*) from card_member;
Query ID = hadoop_20231224055845_aa5d85a4-40b8-4f9c-9dfb-62966c747b54
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1703396126878_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reduce 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 5.99 s
--------------------------------------------------------------------------------
OK
999
Time taken: 6.944 seconds, Fetched: 1 row(s)
hive>
```

Verify the record count in both the tables. These are 999.

```
hive> select * from card_member limit 10;
OK
340028465709212 009250698176266 2012-02-08 06:04:13.0   05/13   United States   Barberton       NULL
340054675199675 835873341185231 2017-03-10 09:24:44.0   03/17   United States   Fort Dodge      NULL
340082915339645 512969555857346 2014-02-15 06:30:30.0   07/14   United States   Graham  NULL
340134186926007 887711945571282 2012-02-05 01:21:58.0   02/13   United States   Dix Hills       NULL
340265728490548 680324265406190 2014-03-29 07:49:14.0   11/14   United States   Rancho Cucamonga        NULL
340268219434811 929799084911715 2012-07-08 02:46:08.0   08/12   United States   San Francisco   NULL
340379737226464 089615510858348 2010-03-10 00:06:42.0   09/10   United States   Clinton NULL
340383645652108 181180599313885 2012-02-24 05:32:44.0   10/16   United States   West New York   NULL
340803866934451 417664728506297 2015-05-21 04:30:45.0   08/17   United States   Beaverton       NULL
340889618969736 459292914761635 2013-04-23 08:40:11.0   11/15   United States   West Palm Beach NULL
Time taken: 0.108 seconds, Fetched: 10 row(s)
```

- **Task 3**: Create a look-up table with columns specified earlier in the problem statement.

8. Create main Lookup table: Added HBASE linkage in case we need to use dao script in future for kafka processing logic.

```
Time taken: 6.944 seconds, Fetched: 1 row(s)
hive> CREATE TABLE card_member_lookup
    > (CARD_ID  STRING,
    > UCL DOUBLE,
    > POSTCODE STRING,
    > TRANSACTION_DT TIMESTAMP,
    > SCORE INT
    > )
    > STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
    > WITH SERDEPROPERTIES ("hbase.columns.mapping"=":key, lookup_card_family:ucl, lookup_card_family:postcode, lookup_card_family:transaction_dt, lookup_card_family:score")
    > TBLPROPERTIES ("hbase.table.name" = "lookup_data_hbase");
OK
Time taken: 2.848 seconds
hive> describe card_member_lookup;
OK
card_id                 string
ucl                     double
postcode                string
transaction_dt          timestamp
score                   int
Time taken: 0.046 seconds, Fetched: 5 row(s)
hive>
```

- **Task 4**: After creating the table, you need to load the relevant data in the lookup table.

9. Load the Lookup table using the command:

```
hive> INSERT OVERWRITE TABLE LOOKUP_DATA_HBASE
    > SELECT trans.card_id,
    >        trans.moving_average+3*standard_deviation as UCL,
    >        POSTCODE,
    >        transaction_dt,
    >        member_score.score
    > FROM
    > (
    > SELECT
    >    card_id,
    >    AVG(amount)
    >          OVER(PARTITION BY card_id ORDER BY transaction_dt ROWS BETWEEN 9 PRECEDING AND CURRENT ROW)
    >          AS moving_average,
    >    STDDEV(amount)
    >          OVER(PARTITION BY card_id ORDER BY transaction_dt ROWS BETWEEN 9 PRECEDING AND CURRENT ROW)
    >          AS standard_deviation,
    >     transaction_dt,
    >     POSTCODE,
    >     ROW_NUMBER() OVER(PARTITION BY card_id ORDER BY transaction_dt DESC ) RN
    > FROM transactions_formatted
    > WHERE STATUS = 'GENUINE'
    > )trans
    > inner JOIN card_member on (trans.card_id=card_member.card_id)
    > inner JOIN member_score on (member_score.MEMBER_ID=card_member.MEMBER_ID)
    > WHERE RN=1;
No Stats for capstone@transactions_formatted, Columns: amount, postcode, transaction_dt, card_id, status
No Stats for capstone@card_member, Columns: member_id, card_id
No Stats for capstone@member_score, Columns: member_id, score
Query ID = hadoop_20231209111801_337a14d4-a8b5-421e-a671-09bd89e6e9e2
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_0011)

--------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 4 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 5 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      2          2        0        0       0       0
--------------------------------------------------------------------------------------------
VERTICES: 05/05  [==========================>>] 100%  ELAPSED TIME: 19.13 s
--------------------------------------------------------------------------------------------
OK
Time taken: 30.255 seconds
```

10. Verify the records in the lookup table:

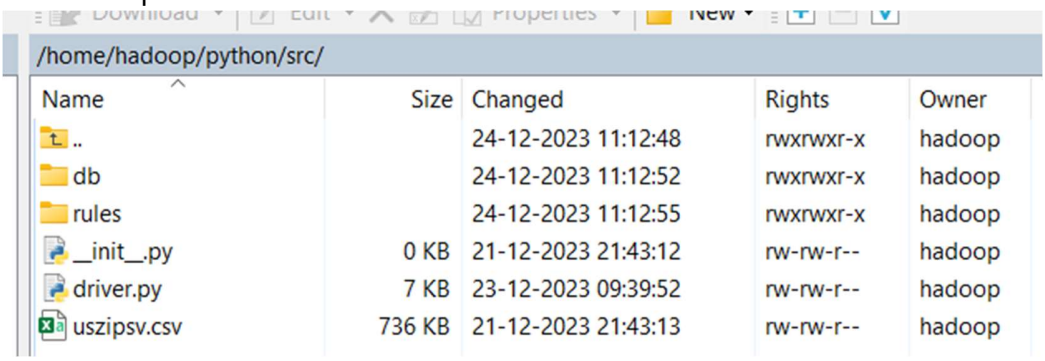
```
Time taken: 21.455 seconds
hive> select * from card_member_lookup;
OK
340028465709212 1.6331555548882348E7    24658   2018-01-02 03:25:35     233
340054675199675 1.4156079786189131E7    50140   2018-01-15 19:43:23     631
340082915339645 1.5285685330791473E7    17844   2018-01-26 19:03:47     407
340134186926007 1.5239767522438556E7    67576   2018-01-18 23:12:50     614
340265728490548 1.608491671255562E7     72435   2018-01-21 02:07:35     202
340268219434811 1.2507323937605347E7    62513   2018-01-16 04:30:05     415
340379737226464 1.4198310998368107E7    26656   2018-01-27 00:19:47     229
```

Logic Final

Following are the steps to be performed for Streaming Kafka for transactions capture:
Make sure you have all the required setup installed like gcc, happybase pandas.
Make sure happybase thrift server is up and running.

```
# to check what processes are running, we check for thrift server
jps
# To install happy base and Pandas setup
yum update -y
yum install gcc
yum install python3-devel
pip install happybase --use-feature=2020-resolver
pip install pandas --use-feature=2020-resolver
```

- Prepare the folder structure as below:



/home/hadoop/python/src/

Name	Size	Changed	Rights	Owner
..		24-12-2023 11:12:48	rwxrwxr-x	hadoop
db		24-12-2023 11:12:52	rwxrwxr-x	hadoop
rules		24-12-2023 11:12:55	rwxrwxr-x	hadoop
__init__.py	0 KB	21-12-2023 21:43:12	rw-rw-r--	hadoop
driver.py	7 KB	23-12-2023 09:39:52	rw-rw-r--	hadoop
uszipsv.csv	736 KB	21-12-2023 21:43:13	rw-rw-r--	hadoop

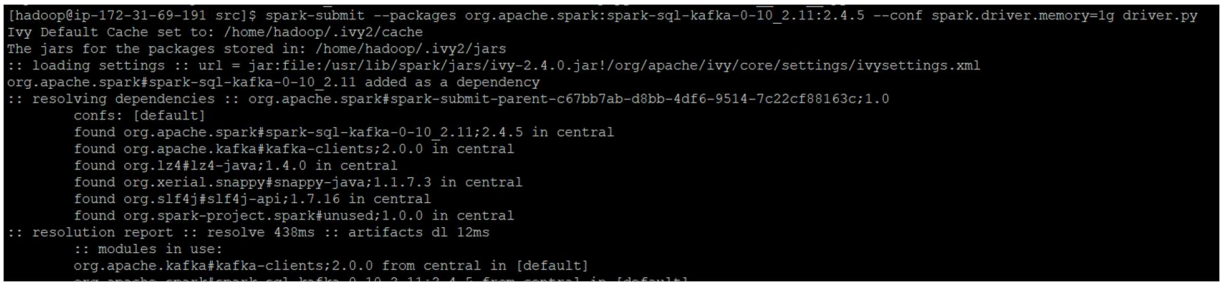
- Place dao and geo_map files under db. Keep rules.py file in rules folder.
- Set Export version in hadoop putty

```
export SPARK_KAFKA_VERSION=0.10
```

- Run the command via cluster

```
spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.
5 --conf spark.driver.memory=1g driver.py
```

Added clause for spark.driver.memory=1g as we were getting not enough resources error.



```
[hadoop@ip-172-31-69-191 src]$ spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5 --conf spark.driver.memory=1g driver.py
Ivy Default Cache set to: /home/hadoop/.ivy2/cache
The jars for the packages stored in: /home/hadoop/.ivy2/jars
:: loading settings :: url = jar:file:/usr/lib/spark/jars/ivy-2.4.0.jar!/org/apache/ivy/core/settings/ivysettings.xml
org.apache.spark#spark-sql-kafka-0-10_2.11 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-c67bb7ab-d8bb-4df6-9514-7c22cf88163c;1.0
        confs: [default]
        found org.apache.spark#spark-sql-kafka-0-10_2.11;2.4.5 in central
        found org.apache.kafka#kafka-clients;2.0.0 in central
        found org.lz4#lz4-java;1.4.0 in central
        found org.xerial.snappy#snappy-java;1.1.7.3 in central
        found org.slf4j#slf4j-api;1.7.16 in central
        found org.spark-project.spark#unused;1.0.0 in central
:: resolution report :: resolve 438ms :: artifacts dl 12ms
        :: modules in use:
        org.apache.kafka#kafka-clients;2.0.0 from central in [default]
```

- After Running the command, wait for some time and Kill the instance using Ctrl+C , as this is supposed to be running all the time however we have to terminate it.

```
23/12/24 07:03:20 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, ip-172-31-69-191.ec2.internal, 43045, None)
23/12/24 07:03:20 INFO BlockManagerMasterEndpoint: Registering block manager ip-172-31-69-191.ec2.internal:43045 with 414.4 MB RAM, BlockManagerId(driver, ip-1
45, None)
23/12/24 07:03:20 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, ip-172-31-69-191.ec2.internal, 43045, None)
23/12/24 07:03:20 INFO BlockManager: external shuffle service port = 7337
23/12/24 07:03:20 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, ip-172-31-69-191.ec2.internal, 43045, None)
23/12/24 07:03:21 INFO EventLoggingListener: Logging events to hdfs:/var/log/spark/apps/local-1703401400000
23/12/24 07:03:21 INFO SharedState: loading hive config file: file:/etc/spark/conf.dist/hive-site.xml
23/12/24 07:03:21 INFO SharedState: Setting hive.metastore.warehouse.dir ('null') to the value of spark.sql.warehouse.dir ('hdfs:///user/spark/warehouse').
23/12/24 07:03:21 INFO SharedState: Warehouse path is 'hdfs:///user/spark/warehouse'.
23/12/24 07:03:22 INFO StateStoreCoordinatorRef: Registered StateStoreCoordinator endpoint
awaiting stream termination
-------------------------------------------
Batch: 0
-------------------------------------------
+----------------+-------------+-------+---------------+--------+-------------------+-------+
|card_id         |member_id    |amount |pos_id         |postcode|transaction_dt_ts  |status |
+----------------+-------------+-------+---------------+--------+-------------------+-------+
|348702330256514 |37495066290  |4380912|2480634068007 22|96774   |2017-12-31 08:24:29|GENUINE|
|348702330256514 |37495066290  |6703385|7865627771408 12|84758   |2017-12-31 04:15:03|GENUINE|
|348702330256514 |37495066290  |7454328|4669525713935 08|93645   |2017-12-31 09:56:42|GENUINE|
|348702330256514 |37495066290  |4013428|4584532033031 9 |15868   |2017-12-31 05:38:54|GENUINE|
|348702330256514 |37495066290  |5495353|5454996219656 97|79033   |2017-12-31 21:51:54|GENUINE|
|348702330256514 |37495066290  |3966214|3692663422725 01|22832   |2017-12-31 03:52:51|GENUINE|
|348702330256514 |37495066290  |1753644|9475029292671   |17923   |2017-12-31 00:11:30|GENUINE|
|348702330256514 |37495066290  |1692115|2764752519586 0 |55708   |2017-12-31 17:02:39|GENUINE|
|5189563368503974|117826301530 |9222134|5257013373551 94|64002   |2017-12-31 20:22:10|GENUINE|
|5189563368503974|117826301530 |4133848|1820313834431 15|26346   |2017-12-31 01:52:32|GENUINE|
|5189563368503974|117826301530 |8938921|7997482464110 19|76934   |2017-12-31 05:20:53|GENUINE|
|5189563368503974|117826301530 |1786366|1312768180712 65|63431   |2017-12-31 14:29:38|GENUINE|
|5189563368503974|117826301530 |9142237|5642402596789 03|50635   |2017-12-31 19:37:19|GENUINE|
|5407073344486464|1147922084344|6885448|8879139067111 17|59031   |2017-12-31 07:53:53|GENUINE|
|5407073344486464|1147922084344|4028209|1162660511181 82|80118   |2017-12-31 01:06:50|GENUINE|
|5407073344486464|1147922084344|3858369|8961058176133 25|53820   |2017-12-31 17:37:26|GENUINE|
|5407073344486464|1147922084344|9307733|7293741160164 79|14898   |2017-12-31 04:50:16|GENUINE|
|5407073344486464|1147922084344|4011296|5433733673196 47|44028   |2017-12-31 13:09:34|GENUINE|
|5407073344486464|1147922084344|9492531|2119800956593 71|49453   |2017-12-31 14:12:26|GENUINE|
|5407073344486464|1147922084344|7550074|3455330881120 99|15030   |2017-12-31 02:34:52|GENUINE|
+----------------+-------------+-------+---------------+--------+-------------------+-------+
only showing top 20 rows
```

- Verifying the records count as below:

```
hive> select count(*) from transactions_formatted;
Query ID = hadoop_20231224070648_2e46f252-bbcd-465a-ba8c-0a5de85ceddc
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1703396126878_0006)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reduce 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 9.03 s
--------------------------------------------------------------------------------
OK
59367
Time taken: 11.984 seconds, Fetched: 1 row(s)
hive>
```