Scripts Execution

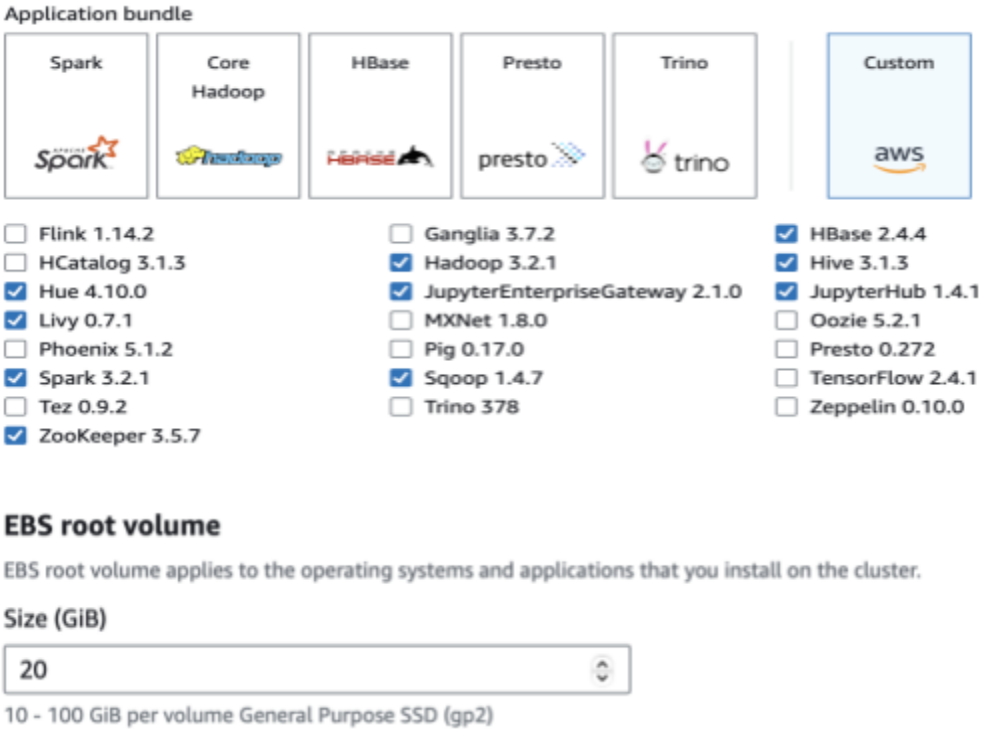
Explanation of the solution to the streaming layer problem.

Step 1: Basic setup for Project as mentioned in brief:

- EMR cluster setup with Hadoop, Sqoop, Hive, HBase and Spark, root device volume size as 20 GB.
- Access EMR cluster using AWS console/ssh.
- Setup the project directory.
- Create a streaming data processing framework which would ingest real time POS transaction data from Kafka and then the transaction data is validated based on three rules parameters (stored in the NoSQL database) as done in the mid-submission.
- Update the transactions data along with the status (fraud/genuine) in the card_transactions table.
- Store the 'postcode' and 'transaction_dt' of the current transaction in the look-up table in the NoSQL database if the transaction was classified as genuine.

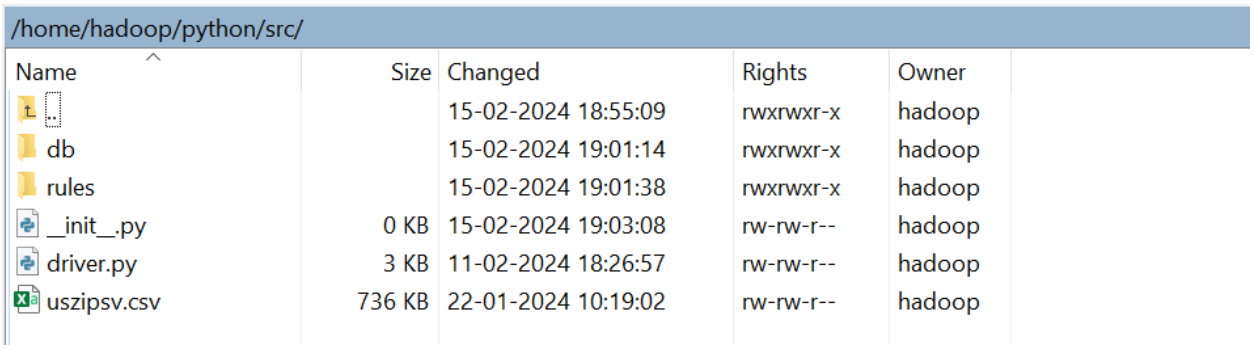
Step1: Setup cluster

1. EMR config



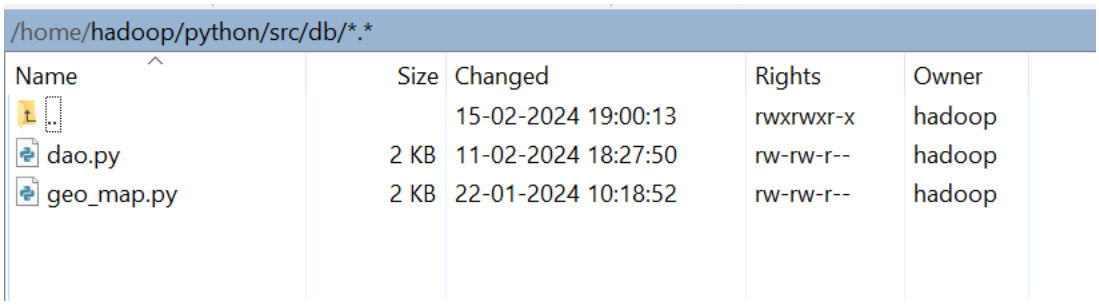
2. A directory named ‘python’ is created with sub-directory named ‘src’.

- ‘src’ directory has two sub-directories ‘db’ and ‘rules’.
- ‘src’ directory also has a python file named driver.p
- ‘rules’ directory has a python file named rules.py and ‘db’ directory has geo_map.py and dao.py
- Downloaded dao.py, geo_map.py, and uszipsv.csv from the resource section of the capstone project via WinSCP to hadoop with python file driver.py



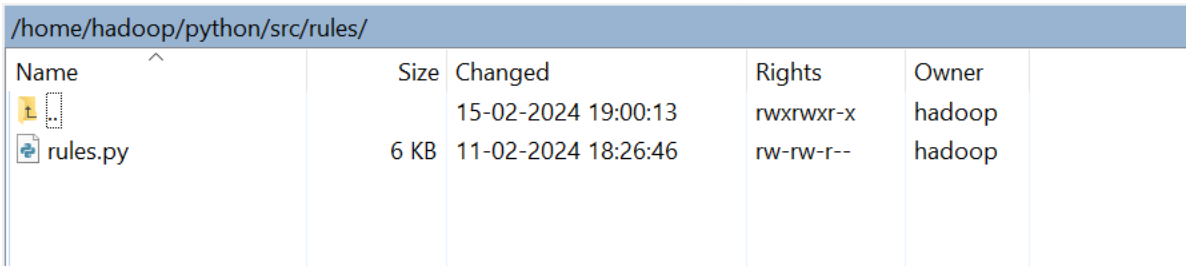
/home/hadoop/python/src/

Name	Size	Changed	Rights	Owner
..		15-02-2024 18:55:09	rwxrwxr-x	hadoop
db		15-02-2024 19:01:14	rwxrwxr-x	hadoop
rules		15-02-2024 19:01:38	rwxrwxr-x	hadoop
__init__.py	0 KB	15-02-2024 19:03:08	rw-rw-r--	hadoop
driver.py	3 KB	11-02-2024 18:26:57	rw-rw-r--	hadoop
uszipsv.csv	736 KB	22-01-2024 10:19:02	rw-rw-r--	hadoop



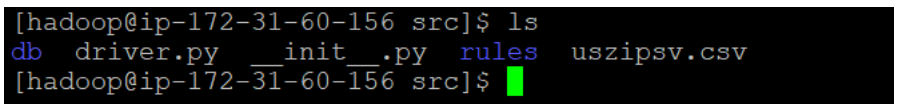
/home/hadoop/python/src/db/*.*

Name	Size	Changed	Rights	Owner
..		15-02-2024 19:00:13	rwxrwxr-x	hadoop
dao.py	2 KB	11-02-2024 18:27:50	rw-rw-r--	hadoop
geo_map.py	2 KB	22-01-2024 10:18:52	rw-rw-r--	hadoop



/home/hadoop/python/src/rules/

Name	Size	Changed	Rights	Owner
..		15-02-2024 19:00:13	rwxrwxr-x	hadoop
rules.py	6 KB	11-02-2024 18:26:46	rw-rw-r--	hadoop



```
[hadoop@ip-172-31-60-156 src]$ ls
db  driver.py  __init__.py  rules  uszipsv.csv
[hadoop@ip-172-31-60-156 src]$
```

3. Switch to root and then Install kafka-python package

```
[root@ip-172-31-55-197 ~]# pip install kafka-python
WARNING: Running pip install with root privileges is generally not a good idea. Try `pip3 install --user` instead.
Collecting kafka-python
  Downloading https://files.pythonhosted.org/packages/75/68/dcb0db055309f680ab2931a3eeb22d865604b638acf8c914bedf4c1a0c8c/kafka_python-2.0.2-py2.py3-none-any.
whl (246kB)
    100% |████████████████████████████████| 256kB 2.1MB/s
Installing collected packages: kafka-python
Successfully installed kafka-python-2.0.2
[root@ip-172-31-55-197 ~]#
```

4. Command to install happybase and start thrift server.
   sudo yum update
   sudo yum install python3-devel
   pip install happybase
   /usr/lib/hbase/bin/hbase-daemon.sh start thrift -p 9090

```
[root@ip-172-31-55-197 ~]# sudo yum update
Loaded plugins: extras_suggestions, langpacks, priorities, update-motd
```

```
  ruby-devel.x86_64 0:2.0.0.648-36.amzn2.0.6                               ruby-irb.noarch 0:2.0.0.648-36.amzn2.0.6
  ruby-libs.x86_64 0:2.0.0.648-36.amzn2.0.6                                rubygem-bigdecimal.x86_64 0:1.2.0-36.amzn2.0.6
  rubygem-io-console.x86_64 0:0.4.2-36.amzn2.0.6                           rubygem-json.x86_64 0:1.7.7-36.amzn2.0.6
  rubygem-psych.x86_64 0:2.0.0-36.amzn2.0.6                                rubygem-rdoc.noarch 0:4.0.0-36.amzn2.0.6
  rubygems.noarch 0:2.0.14.1-36.amzn2.0.6                                  selinux-policy.noarch 0:3.13.1-192.amzn2.6.8
  selinux-policy-targeted.noarch 0:3.13.1-192.amzn2.6.8                    shadow-utils.x86_64 2:4.1.5.1-24.amzn2.0.3
  strace.x86_64 0:4.26-1.amzn2.0.1                                         sysctl-defaults.noarch 0:1.0-3.amzn2
  system-lsb.x86_64 0:4.1-27.amzn2.3.6                                     system-lsb-core.x86_64 0:4.1-27.amzn2.3.6
  system-lsb-cxx.x86_64 0:4.1-27.amzn2.3.6                                 system-lsb-desktop.x86_64 0:4.1-27.amzn2.3.6
  system-lsb-languages.x86_64 0:4.1-27.amzn2.3.6                           system-lsb-submod-multimedia.x86_64 0:4.1-27.amzn2.3.6
  system-lsb-submod-security.x86_64 0:4.1-27.amzn2.3.6                     system-release.x86_64 1:2-16.amzn2
  systemtap-runtime.x86_64 0:4.5-1.amzn2.0.1                               systemtap-sdt-devel.x86_64 0:4.5-1.amzn2.0.1
  tcpdump.x86_64 14:4.9.2-4.amzn2.1.0.1                                    texlive-base.noarch 2:2012-38.20130427_r30134.amzn2.0.5
  texlive-dvipng.noarch 2:svn26689.1.14-38.amzn2.0.5                       texlive-kpathsea.noarch 2:svn28792.0-38.amzn2.0.5
  traceroute.x86_64 3:2.0.22-2.amzn2.0.2                                   tzdata.noarch 0:2023d-1.amzn2.0.1
  update-motd.noarch 0:1.1.2-2.amzn2.0.2                                   xdg-utils.noarch 0:1.1.0-0.17.20120809git.amzn2.0.1
  xfsprogs.x86_64 0:5.0.0-10.amzn2.0.1                                     yajl.x86_64 0:2.0.4-4.amzn2.0.3

Replaced:
  python-colorama.noarch 0:0.3.2-3.amzn2          python-six.noarch 0:1.9.0-2.amzn2          system-lsb-printing.x86_64 0:4.1-27.amzn2.3.5

Complete!
```

```
[root@ip-172-31-55-197 ~]# sudo yum install python3-devel
Loaded plugins: extras_suggestions, langpacks, priorities, update-motd
Existing lock /var/run/yum.pid: another copy is running as pid 25098.
Another app is currently holding the yum lock; waiting for it to exit...
  The other application is: yum
    Memory : 145 M RSS (450 MB VSZ)
    Started: Sun Feb 18 12:54:34 2024 - 00:22 ago
    State  : Sleeping, pid: 25098
14 packages excluded due to repository priority protections
Resolving Dependencies
--> Running transaction check
---> Package python3-devel.x86_64 0:3.7.16-1.amzn2.0.4 will be installed
--> Processing Dependency: python3-rpm-macros for package: python3-devel-3.7.16-1.amzn2.0.4.x86_64
--> Running transaction check
---> Package python3-rpm-macros.noarch 0:3-60.amzn2.0.1 will be installed
--> Processing Dependency: python-srpm-macros >= 3-38 for package: python3-rpm-macros-3-60.amzn2.0.1.noarch
--> Processing Dependency: python-rpm-macros for package: python3-rpm-macros-3-60.amzn2.0.1.noarch
--> Running transaction check
---> Package python-rpm-macros.noarch 0:3-60.amzn2.0.1 will be installed
---> Package python-srpm-macros.noarch 0:3-60.amzn2.0.1 will be installed
--> Finished Dependency Resolution

Dependencies Resolved

================================================================================================================
 Package                   Arch               Version                          Repository                  Size
================================================================================================================
Installing:
 python3-devel             x86_64             3.7.16-1.amzn2.0.4               amzn2-core                 244 k
Installing for dependencies:
 python-rpm-macros         noarch             3-60.amzn2.0.1                   amzn2-core                  14 k
 python-srpm-macros        noarch             3-60.amzn2.0.1                   amzn2-core                  18 k
 python3-rpm-macros        noarch             3-60.amzn2.0.1                   amzn2-core                  12 k

Transaction Summary
================================================================================================================
Install  1 Package (+3 Dependent packages)

Total download size: 289 k
Installed size: 753 k
Is this ok [y/d/N]: y
Downloading packages:
(1/4): python-srpm-macros-3-60.amzn2.0.1.noarch.rpm                                    |  18 kB  00:00:00
(2/4): python-rpm-macros-3-60.amzn2.0.1.noarch.rpm                                     |  14 kB  00:00:00
```

```
Installed size: 753 k
Is this ok [y/d/N]: y
Downloading packages:
(1/4): python-srpm-macros-3-60.amzn2.0.1.noarch.rpm                                                 |  18 kB  00:00:00
(2/4): python-rpm-macros-3-60.amzn2.0.1.noarch.rpm                                                  |  14 kB  00:00:00
(3/4): python3-rpm-macros-3-60.amzn2.0.1.noarch.rpm                                                 |  12 kB  00:00:00
(4/4): python3-devel-3.7.16-1.amzn2.0.4.x86_64.rpm                                                  | 244 kB  00:00:00
------------------------------------------------------------------------------------------------------------------------
Total                                                                              2.6 MB/s | 289 kB  00:00:00
Running transaction check
Running transaction test
Transaction test succeeded
Running transaction
  Installing : python-srpm-macros-3-60.amzn2.0.1.noarch                                                         1/4
  Installing : python-rpm-macros-3-60.amzn2.0.1.noarch                                                          2/4
  Installing : python3-rpm-macros-3-60.amzn2.0.1.noarch                                                         3/4
  Installing : python3-devel-3.7.16-1.amzn2.0.4.x86_64                                                          4/4
  Verifying  : python-rpm-macros-3-60.amzn2.0.1.noarch                                                          1/4
  Verifying  : python-srpm-macros-3-60.amzn2.0.1.noarch                                                         2/4
  Verifying  : python3-rpm-macros-3-60.amzn2.0.1.noarch                                                         3/4
  Verifying  : python3-devel-3.7.16-1.amzn2.0.4.x86_64                                                          4/4

Installed:
  python3-devel.x86_64 0:3.7.16-1.amzn2.0.4

Dependency Installed:
  python-rpm-macros.noarch 0:3-60.amzn2.0.1          python-srpm-macros.noarch 0:3-60.amzn2.0.1          python3-rpm-macros.noarch 0:3-60.amzn2.0.1

Complete!
```

```
[root@ip-172-31-55-197 ~]# pip install happybase
WARNING: Running pip install with root privileges is generally not a good idea. Try `pip3 install --user` instead.

Collecting happybase
  Downloading happybase-1.2.0.tar.gz (40 kB)
     |████████████████████████████████| 40 kB 6.4 MB/s
Requirement already satisfied: six in /usr/local/lib/python3.7/site-packages (from happybase) (1.13.0)
Collecting thriftpy2>=0.4
  Downloading thriftpy2-0.4.17.tar.gz (519 kB)
     |████████████████████████████████| 519 kB 35.3 MB/s
  Installing build dependencies ... done
  WARNING: Missing build requirements in pyproject.toml for thriftpy2>=0.4 from https://files.pythonhosted.org/packages/1d/5c/852a627317a75e0ec19f42b955ef115
b0906c43ee4c7595c112a652f0b20/thriftpy2-0.4.17.tar.gz#sha256=190f35c32da9146d1fdd822f46b6a0ad543572ea405ca6853b4ec7b128efbc0d (from happybase).
  WARNING: The project does not specify a build backend, and pip cannot fall back to setuptools without 'wheel'.
  Getting requirements to build wheel ... done
  Installing backend dependencies ... done
    Preparing wheel metadata ... done
Collecting ply<4.0,>=3.4
  Downloading ply-3.11-py2.py3-none-any.whl (49 kB)
     |████████████████████████████████| 49 kB 11.2 MB/s
Using legacy 'setup.py install' for happybase, since package 'wheel' is not installed.
Building wheels for collected packages: thriftpy2
  Building wheel for thriftpy2 (PEP 517) ... done
  Created wheel for thriftpy2: filename=thriftpy2-0.4.17-cp37-cp37m-linux_x86_64.whl size=1308555 sha256=b6f67e93640afa4613b38bd406a30b5d72635aab6d463e90413d
766206f265cd
  Stored in directory: /root/.cache/pip/wheels/6a/f3/2b/9a4b02cc3cdff27f9afc9101d3df6de13e975caef66c7dcb77
Successfully built thriftpy2
Installing collected packages: ply, thriftpy2, happybase
    Running setup.py install for happybase ... done
ERROR: After October 2020 you may experience errors when installing or updating packages. This is because pip will change the way that it resolves dependency
 conflicts.

We recommend you use --use-feature=2020-resolver to test your packages with the new resolver before it becomes the default.

thriftpy2 0.4.17 requires six~=1.15, but you'll have six 1.13.0 which is incompatible.
Successfully installed happybase-1.2.0 ply-3.11 thriftpy2-0.4.17
```

```
[root@ip-172-31-55-197 ~]# /usr/lib/hbase/bin/hbase-daemon.sh start thrift -p 9090
running thrift, logging to /usr/lib/hbase/bin/../logs/hbase-root-thrift-ip-172-31-55-197.out
```

5. Command to check if happybase has been installed successfully.

```
[root@ip-172-31-55-197 ~]# python -c "import happybase"
[root@ip-172-31-55-197 ~]#
```

6. We have updated the public IP of EC2 instance “44.192.121.236“ self.host in dao.py file.

```python
import happybase

class HBaseDao:
    """
    Dao class for operation on HBase
    """
    __instance = None

    @staticmethod
    def get_instance():
        """ Static access method. """
        if HBaseDao.__instance == None:
            HBaseDao()
        return HBaseDao.__instance

    def __init__(self):
        if HBaseDao.__instance != None:
            raise Exception("This class is a singleton!")
        else:
            HBaseDao.__instance = self
            self.host = '44.192.121.236'
```

7. Add table details in rules.py.

```python
#Importing all the libraries
from db.dao import HBaseDao
from db.geo_map import GEO_Map
from datetime import datetime
import uuid

lookup_table = 'lookup_data_hive'
card_trans_table = 'card_transactions_hive'
```

8. Python function for the three given rules.
    a. **First rule (verify_ucl_data):** Function to verify transaction as genuine if transaction amount is less than Upper control limit (UCL).

```python
class Rules():

    def verify_ucl_data(card_id, amount):
        try:
            hbasedao = HBaseDao.get_instance()
            card_row = hbasedao.get_data(key=str(card_id), table=lookup_table)
            card_ucl = (card_row[b'lookup_card_family:ucl']).decode("utf-8")

            if amount < float(card_ucl):
                return True
            else:
                return False
        except Exception as e:
            raise Exception(e)
```

    b. **Second rule (verify_credit_score_data):** Function to verify transaction as genuine if credit score is greated than 200.

```
def verify_credit_score_data(card_id):

    try:
        hbasedao = HBaseDao.get_instance()

        card_row = hbasedao.get_data(key=str(card_id), table=lookup_table)
        card_score = (card_row[b'lookup_card_family:score']).decode("utf-8")

        if int(card_score) > 200:
            return True
        else:
            return False
    except Exception as e:
        raise Exception(e)
```

c. **Third rule (verify_postcode_data) and (calculate_speed) :** Function to verify transaction as genuine if distance between the current transaction and the last transaction location with respect to time and zipcode is less than a particular threshold.

```
def verify_postcode_data(card_id, postcode, transaction_dt):
    try:
        hbasedao = HBaseDao.get_instance()
        geo_map = GEO_Map.get_instance()

        card_row = hbasedao.get_data(key=str(card_id), table=lookup_table)
        last_postcode = (card_row[b'lookup_transaction_family:postcode']).decode("utf-8")
        last_transaction_dt = (card_row[b'lookup_transaction_family:transaction_dt']).decode("utf-8")

        current_lat = geo_map.get_lat(str(postcode))
        current_lon = geo_map.get_long(str(postcode))
        previous_lat = geo_map.get_lat(last_postcode)
        previous_lon = geo_map.get_long(last_postcode)

        dist = geo_map.distance(lat1=current_lat, long1=current_lon, lat2=previous_lat, long2=previous_lon)

        speed = calculate_speed(dist, transaction_dt, last_transaction_dt)

        if speed < speed_threshold:
            return True
        else:
            return False

    except Exception as e:
        raise Exception(e)
```

```
def calculate_speed(dist, transaction_dt1, transaction_dt2):
    transaction_dt1 = datetime.strptime(transaction_dt1, '%d-%m-%Y %H:%M:%S')
    transaction_dt2 = datetime.strptime(transaction_dt2, '%d-%m-%Y %H:%M:%S')

    elapsed_time = transaction_dt1 - transaction_dt2
    elapsed_time = elapsed_time.total_seconds()

    try:
        return dist / elapsed_time
    except ZeroDivisionError:
        return 299792.458
```

d. Function (**verify_rules_status**): A function to verify all the three above rules – UCL, Credit score and speed.

```
def verify_rules_status(card_id, member_id, amount, pos_id, postcode, transaction_dt):

    hbasedao = HBaseDao.get_instance()

    rule1 = verify_ucl_data(card_id, amount)
    rule2 = verify_credit_score_data(card_id)
    rule3 = verify_postcode_data(card_id, postcode, transaction_dt)

    if all([rule1, rule2, rule3]):
        status = 'GENUINE'
        hbasedao.write_data(key=str(card_id),
                            row={'lookup_transaction_family:postcode': str(postcode), 'lookup_transaction_family:transaction_dt': str(transaction_dt)},
                            table=lookup_table)
    else:
        status = 'FRAUD'

    new_id = str(uuid.uuid4()).replace('-', '')
    hbasedao.write_data(key=new_id,
                        row={'card_transactions_family:card_id': str(card_id), 'card_transactions_family:member_id': str(member_id),
                             'card_transactions_family:amount': str(amount), 'card_transactions_family:pos_id': str(pos_id),
                             'card_transactions_family:postcode': str(postcode), 'card_transactions_family:status': str(status),
                             'card_transactions_family:transaction_dt': str(transaction_dt)},
                        table=card_trans_table)
    return status
```

9. Next, we imported required libraries and modules. Initializing spark session and setting up configuration to connect to Kafka.
   Bootstrap-server: 18.211.252.152
   Port Number: 9092
   Topic: transactions-topic-verified

```
import os
import sys
from pyspark.sql import SparkSession
from pyspark.sql.functions import *
from pyspark.sql.types import *
from rules.rules import Rules

# Initialising SparkSession
spark = SparkSession  \
    .builder  \
    .appName("CreditCardFraudDetection")  \
    .getOrCreate()
spark.sparkContext.setLogLevel('ERROR')

# Reading data from Kafka
credit_data = spark.readStream \
    .format("kafka") \
    .option("kafka.bootstrap.servers", "18.211.252.152:9092") \
    .option("startingOffsets","earliest") \
    .option("failOnDataLoss", "false") \
    .option("subscribe", "transactions-topic-verified") \
    .load()
```

10. Define transaction schema in form of JSON.

```
# Defining schema for transaction
dataSchema = StructType() \
    .add("card_id", LongType()) \
    .add("member_id", LongType()) \
    .add("amount", DoubleType()) \
    .add("pos_id", LongType()) \
    .add("postcode", IntegerType()) \
    .add("transaction_dt", StringType())
```

11. Read JSON data from Kafka and validate all the 3 rules defined above.

```
credit_data = credit_data.selectExpr("cast(value as string)")
credit_data_stream = credit_data.select(from_json(col="value", schema=dataSchema).alias("credit_data")).select(
    "credit_data.*")

# Define UDF which verifies all the rules for each transaction and updates the lookup and card transactions tables
verify_all_rules = udf(Rules.verify_rules_status, StringType())

final_data = credit_data_stream \
    .withColumn('status', verify_all_rules(credit_data_stream['card_id'],
                                           credit_data_stream['member_id'],
                                           credit_data_stream['amount'],
                                           credit_data_stream['pos_id'],
                                           credit_data_stream['postcode'],
                                           credit_data_stream['transaction_dt']))
```

12. Code for Displaying output in console and spark termination..

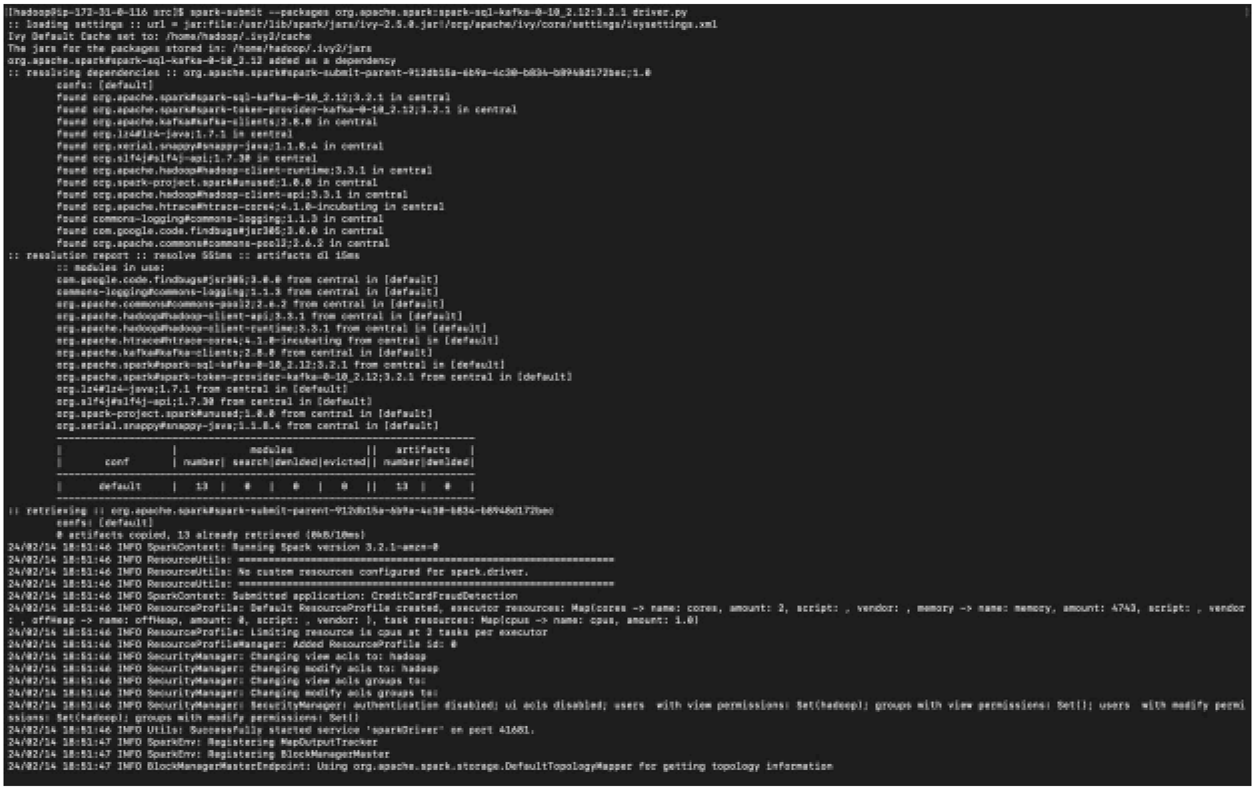
```
# Writes output to console
output_data = final_data \
    .select("card_id", "member_id", "amount", "pos_id", "postcode", "transaction_dt") \
    .writeStream \
    .outputMode("append") \
    .format("console") \
    .option("truncate", False) \
    .start()

# Indicating Spark to await termination
output_data.awaitTermination()
```

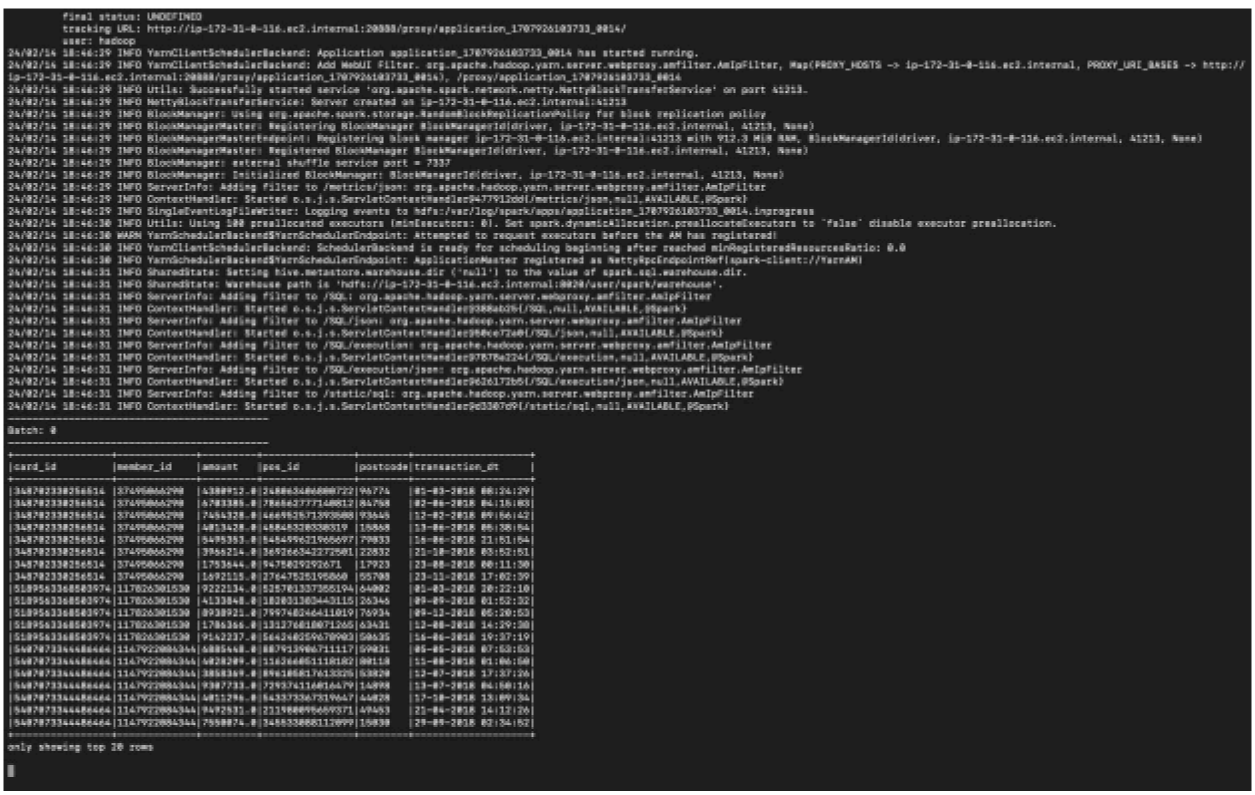
13. Setting up Kafka version and to run the spark-submit command.

```
[hadoop@ip-172-31-0-116 src]$ export SPARK_KAFKA_VERSION=0.10
[hadoop@ip-172-31-0-116 src]$ spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.12:3.2.1 driver.py
```

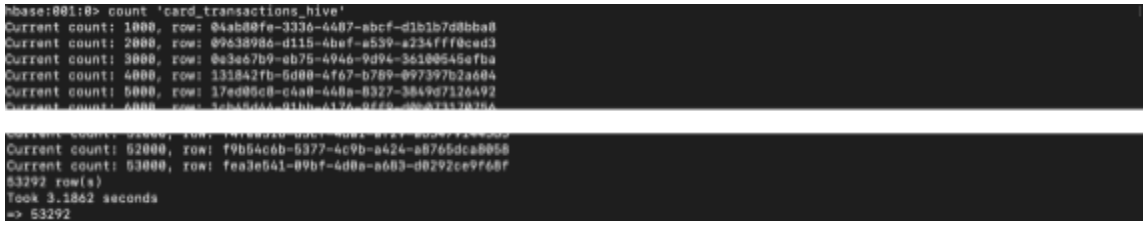
14. Console output

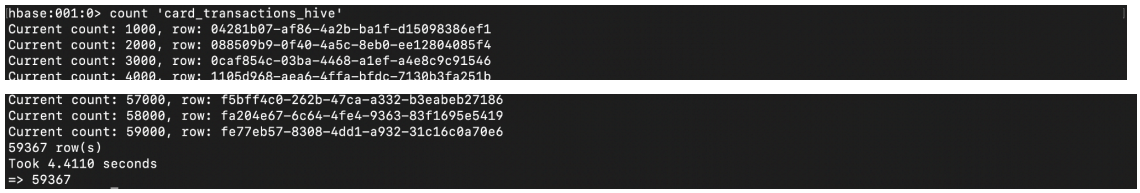


```
[hadoop@ip-172-31-0-116 src]$ spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.12:3.2.1 driver.py
:: loading settings :: url = jar:file:/usr/lib/spark/jars/ivy-2.5.0.jar!/org/apache/ivy/core/settings/ivysettings.xml
Ivy Default Cache set to: /home/hadoop/.ivy2/cache
The jars for the packages stored in: /home/hadoop/.ivy2/jars
org.apache.spark#spark-sql-kafka-0-10_2.12 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-912db15a-6b9a-4c30-b834-b0948d172bec;1.0
        confs: [default]
        found org.apache.spark#spark-sql-kafka-0-10_2.12;3.2.1 in central
        found org.apache.spark#spark-token-provider-kafka-0-10_2.12;3.2.1 in central
        found org.apache.kafka#kafka-clients;2.8.0 in central
        found org.lz4#lz4-java;1.7.1 in central
        found org.xerial.snappy#snappy-java;1.1.8.4 in central
        found org.slf4j#slf4j-api;1.7.30 in central
        found org.apache.hadoop#hadoop-client-runtime;3.3.1 in central
        found org.spark-project.spark#unused;1.0.0 in central
        found org.apache.hadoop#hadoop-client-api;3.3.1 in central
        found org.apache.htrace#htrace-core4;4.1.0-incubating in central
        found commons-logging#commons-logging;1.1.3 in central
        found com.google.code.findbugs#jsr305;3.0.0 in central
        found org.apache.commons#commons-pool2;2.6.2 in central
:: resolution report :: resolve 551ms :: artifacts dl 16ms
        :: modules in use:
        com.google.code.findbugs#jsr305;3.0.0 from central in [default]
        commons-logging#commons-logging;1.1.3 from central in [default]
        org.apache.commons#commons-pool2;2.6.2 from central in [default]
        org.apache.hadoop#hadoop-client-api;3.3.1 from central in [default]
        org.apache.hadoop#hadoop-client-runtime;3.3.1 from central in [default]
        org.apache.htrace#htrace-core4;4.1.0-incubating from central in [default]
        org.apache.kafka#kafka-clients;2.8.0 from central in [default]
        org.apache.spark#spark-sql-kafka-0-10_2.12;3.2.1 from central in [default]
        org.apache.spark#spark-token-provider-kafka-0-10_2.12;3.2.1 from central in [default]
        org.lz4#lz4-java;1.7.1 from central in [default]
        org.slf4j#slf4j-api;1.7.30 from central in [default]
        org.spark-project.spark#unused;1.0.0 from central in [default]
        org.xerial.snappy#snappy-java;1.1.8.4 from central in [default]
        ---------------------------------------------------------------------
        |                  |            modules            ||   artifacts   |
        |       conf       | number| search|dwnlded|evicted|| number|dwnlded|
        ---------------------------------------------------------------------
        |      default     |   13  |   0   |   0   |   0   ||   13  |   0   |
        ---------------------------------------------------------------------
:: retrieving :: org.apache.spark#spark-submit-parent-912db15a-6b9a-4c30-b834-b0948d172bec
        confs: [default]
        0 artifacts copied, 13 already retrieved (0kB/10ms)
24/02/14 18:51:46 INFO SparkContext: Running Spark version 3.2.1-amzn-0
24/02/14 18:51:46 INFO ResourceUtils: ==============================================================
24/02/14 18:51:46 INFO ResourceUtils: No custom resources configured for spark.driver.
24/02/14 18:51:46 INFO ResourceUtils: ==============================================================
24/02/14 18:51:46 INFO SparkContext: Submitted application: CreditCardFraudDetection
24/02/14 18:51:46 INFO ResourceProfile: Default ResourceProfile created, executor resources: Map(cores -> name: cores, amount: 2, script: , vendor: , memory -> name: memory, amount: 4743, script: , vendor
: , offHeap -> name: offHeap, amount: 0, script: , vendor: ), task resources: Map(cpus -> name: cpus, amount: 1.0)
24/02/14 18:51:46 INFO ResourceProfile: Limiting resource is cpus at 2 tasks per executor
24/02/14 18:51:46 INFO ResourceProfileManager: Added ResourceProfile id: 0
24/02/14 18:51:46 INFO SecurityManager: Changing view acls to: hadoop
24/02/14 18:51:46 INFO SecurityManager: Changing modify acls to: hadoop
24/02/14 18:51:46 INFO SecurityManager: Changing view acls groups to:
24/02/14 18:51:46 INFO SecurityManager: Changing modify acls groups to:
24/02/14 18:51:46 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(hadoop); groups with view permissions: Set(); users  with modify permi
ssions: Set(hadoop); groups with modify permissions: Set()
24/02/14 18:51:46 INFO Utils: Successfully started service 'sparkDriver' on port 41681.
24/02/14 18:51:47 INFO SparkEnv: Registering MapOutputTracker
24/02/14 18:51:47 INFO SparkEnv: Registering BlockManagerMaster
24/02/14 18:51:47 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
```



```
         final status: UNDEFINED
         tracking URL: http://ip-172-31-0-116.ec2.internal:20888/proxy/application_1707926103733_0014/
         user: hadoop
24/02/14 18:46:29 INFO YarnClientSchedulerBackend: Application application_1707926103733_0014 has started running.
24/02/14 18:46:29 INFO YarnClientSchedulerBackend: Add WebUI Filter. org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter, Map(PROXY_HOSTS -> ip-172-31-0-116.ec2.internal, PROXY_URI_BASES -> http://
ip-172-31-0-116.ec2.internal:20888/proxy/application_1707926103733_0014), /proxy/application_1707926103733_0014
24/02/14 18:46:29 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 41213.
24/02/14 18:46:29 INFO NettyBlockTransferService: Server created on ip-172-31-0-116.ec2.internal:41213
24/02/14 18:46:29 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
24/02/14 18:46:29 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, ip-172-31-0-116.ec2.internal, 41213, None)
24/02/14 18:46:29 INFO BlockManagerMasterEndpoint: Registering block manager ip-172-31-0-116.ec2.internal:41213 with 912.3 MiB RAM, BlockManagerId(driver, ip-172-31-0-116.ec2.internal, 41213, None)
24/02/14 18:46:29 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, ip-172-31-0-116.ec2.internal, 41213, None)
24/02/14 18:46:29 INFO BlockManager: external shuffle service port = 7337
24/02/14 18:46:29 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, ip-172-31-0-116.ec2.internal, 41213, None)
24/02/14 18:46:29 INFO ServerInfo: Adding filter to /metrics/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/02/14 18:46:29 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@477912dd{/metrics/json,null,AVAILABLE,@Spark}
24/02/14 18:46:29 INFO SingleEventLogFileWriter: Logging events to hdfs:/var/log/spark/apps/application_1707926103733_0014.inprogress
24/02/14 18:46:30 INFO Utils: Using 100 preallocated executors (minExecutors: 0). Set spark.dynamicAllocation.preallocateExecutors to `false` disable executor preallocation.
24/02/14 18:46:30 WARN YarnSchedulerBackend$YarnSchedulerEndpoint: Attempted to request executors before the AM has registered!
24/02/14 18:46:30 INFO YarnClientSchedulerBackend: SchedulerBackend is ready for scheduling beginning after reached minRegisteredResourcesRatio: 0.0
24/02/14 18:46:30 INFO YarnSchedulerBackend$YarnSchedulerEndpoint: ApplicationMaster registered as NettyRpcEndpointRef(spark-client://YarnAM)
24/02/14 18:46:31 INFO SharedState: Setting hive.metastore.warehouse.dir ('null') to the value of spark.sql.warehouse.dir.
24/02/14 18:46:31 INFO SharedState: Warehouse path is 'hdfs://ip-172-31-0-116.ec2.internal:8020/user/spark/warehouse'.
24/02/14 18:46:31 INFO ServerInfo: Adding filter to /SQL: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/02/14 18:46:31 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@388ab29{/SQL,null,AVAILABLE,@Spark}
24/02/14 18:46:31 INFO ServerInfo: Adding filter to /SQL/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/02/14 18:46:31 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@90ce72a0{/SQL/json,null,AVAILABLE,@Spark}
24/02/14 18:46:31 INFO ServerInfo: Adding filter to /SQL/execution: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/02/14 18:46:31 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@7878a224{/SQL/execution,null,AVAILABLE,@Spark}
24/02/14 18:46:31 INFO ServerInfo: Adding filter to /SQL/execution/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/02/14 18:46:31 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@626172b5{/SQL/execution/json,null,AVAILABLE,@Spark}
24/02/14 18:46:31 INFO ServerInfo: Adding filter to /static/sql: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/02/14 18:46:31 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@d3307d9{/static/sql,null,AVAILABLE,@Spark}
-------------------------------------------
Batch: 0
-------------------------------------------
+----------------+-------------+---------+---------------+--------+-------------------+
|card_id         |member_id    |amount   |pos_id         |postcode|transaction_dt     |
+----------------+-------------+---------+---------------+--------+-------------------+
|348702330256514 |37495066290  |4380912.0|248063406800722|96774   |01-03-2018 08:24:29|
|348702330256514 |37495066290  |6703385.0|786562777140812|84758   |02-06-2018 04:15:03|
|348702330256514 |37495066290  |7454328.0|466952571393508|93645   |12-02-2018 09:56:42|
|348702330256514 |37495066290  |4013428.0|45845320330319 |15868   |13-06-2018 05:38:54|
|348702330256514 |37495066290  |5495353.0|545499621965697|79033   |16-06-2018 21:51:54|
|348702330256514 |37495066290  |3966214.0|369266342272501|22832   |21-10-2018 03:52:51|
|348702330256514 |37495066290  |1753644.0|9475029292671  |17923   |23-08-2018 00:11:30|
|348702330256514 |37495066290  |1692115.0|27647525195860 |55708   |23-11-2018 17:02:39|
|5189563368503974|117826301530 |9222134.0|525701337355194|64002   |01-03-2018 20:22:10|
|5189563368503974|117826301530 |4133848.0|182031383443115|26346   |09-09-2018 01:52:32|
|5189563368503974|117826301530 |8938921.0|799748246411019|76934   |09-12-2018 05:20:53|
|5189563368503974|117826301530 |1786366.0|131276818071265|63431   |12-08-2018 14:29:38|
|5189563368503974|117826301530 |9142237.0|564240259678903|50635   |16-06-2018 19:37:19|
|5407073344486464|1147922084344|6885448.0|887913906711117|59031   |05-05-2018 07:53:53|
|5407073344486464|1147922084344|4028209.0|116266051118182|80118   |11-08-2018 01:06:50|
|5407073344486464|1147922084344|3858369.0|896105817613325|53820   |12-07-2018 17:37:26|
|5407073344486464|1147922084344|9307733.0|729374116016479|14898   |13-07-2018 04:50:16|
|5407073344486464|1147922084344|4011296.0|543373367319647|44028   |17-10-2018 13:09:34|
|5407073344486464|1147922084344|9492531.0|211980095659371|49453   |21-04-2018 14:12:26|
|5407073344486464|1147922084344|7550074.0|345533088112099|15030   |29-09-2018 02:34:52|
+----------------+-------------+---------+---------------+--------+-------------------+
only showing top 20 rows
```

15. Count of rows in lookup_data_hive
    a. Before streaming
        i.


    b. After streaming
        i.


        ii.