SKaMP. Tests

The goal of this report is to provide information on the performed testing of data acquisition, data pre-processing, batch processing and streaming processing.

GitHub repository: https://github.com/salveendutt/Big-Data-Analytics.

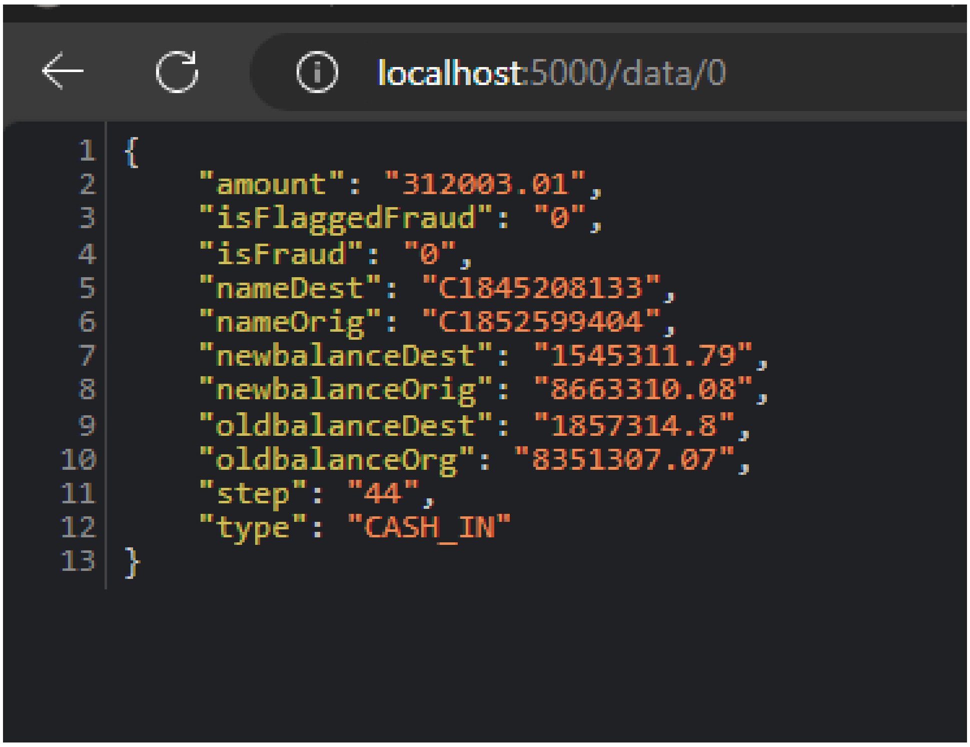
1 Test Scenarios

Test objective	Steps	Expected Result	Actual Result
Verify data incoming from stream API	1. Start the server using start_containers.bat; 2. Navigate to http://localhost:5000	Incoming data is available on /data/0	Passed. The screenshot is provided in Figure 1
Verify correct setup of the stream and data preprocessing functions	Run 'pytest' from the root folder	Data stream is configured as expected; Incoming data is not null; Returned status code - 200. Preprocessing utils return transformed data as expected	Passed. The screenshot is provided in Figure 10
Verify the correct setup of Nifi - HDFS/Kafka flow	Run the containers - follow steps in README.md	Data flows from streamin API to Kafka topics and Hive tables	Passed. The screenshot is provided in Figure 2, 3, 4, 5
Verify the correct setup of batch processing	Run the containers - follow steps in README.md	Views are available in the Cassandra tables	Passed. The screenshot is provided in Figure 6, 7, 8
Verify the correct setup of streaming processing	Run the containers - follow steps in README.md	Views are available in the Cassandra tables	Passed. The screenshot is provided in Figure 9

Table 1: Test scenarios

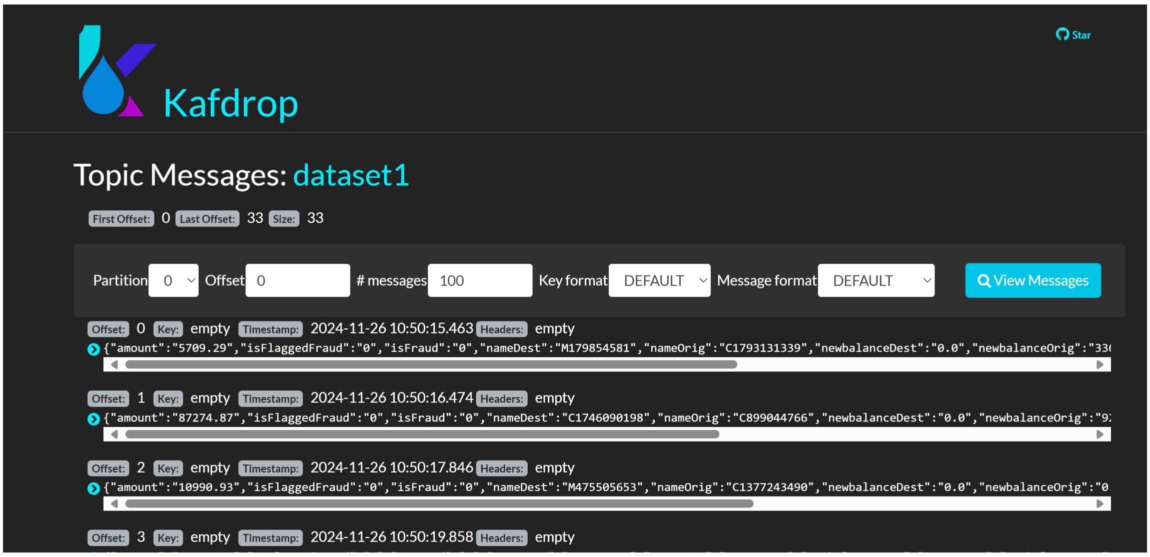
Test objective	Steps	Expected Result	Actual Result
Verify correct data preprocessing of dataset 1	Run ’pytest’ from the root folder	Feature ’type’ is correctly transformed into numeric value (5 cases); Feature ’isMerchant’ is correctly prepared (2 cases)	PASSED. The screenshot is provided in Figure 10
Verify correct data preprocessing of dataset 2	Run ’pytest’ from the root folder	Numeric boolean values are transformed to int from float (4 cases)	PASSED. The screenshot is provided in Figure 10
Verify correct data preprocessing of dataset 3	Run ’pytest’ from the root folder	Feature ’entry_mode’ is correctly transformed into numeric value (4 cases); Unnecessary features are omitted.	PASSED. The screenshot is provided in Figure 10
Verify correct data preprocessing of dataset 4	Run ’pytest’ from the root folder	Features ’Amount’, ’Class’ are renamed to ’amount’ and ’isFraud’; Extra features are removed	PASSED. The screenshot is provided in Figure 10

Table 2: Data pre-processing tests



```
{
    "amount": "312003.01",
    "isFlaggedFraud": "0",
    "isFraud": "0",
    "nameDest": "C1845208133",
    "nameOrig": "C1852599404",
    "newbalanceDest": "1545311.79",
    "newbalanceOrig": "8663310.08",
    "oldbalanceDest": "1857314.8",
    "oldbalanceOrg": "8351307.07",
    "step": "44",
    "type": "CASH_IN"
}
```

Figure 1: Data incoming via the stream



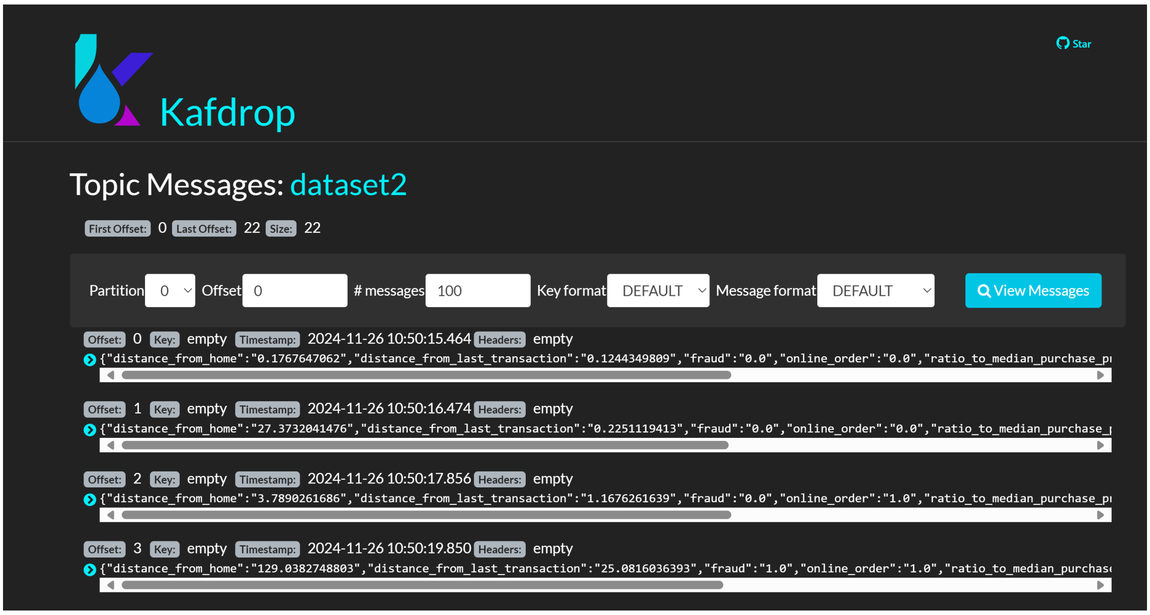
Topic Messages: dataset1

First Offset: 0 Last Offset: 33 Size: 33

Partition 0 Offset 0 # messages 100 Key format DEFAULT Message format DEFAULT

```
Offset: 0 Key: empty Timestamp: 2024-11-26 10:50:15.463 Headers: empty
{"amount":"5709.29","isFlaggedFraud":"0","isFraud":"0","nameDest":"M179854581","nameOrig":"C1793131339","newbalanceDest":"0.0","newbalanceOrig":"33

Offset: 1 Key: empty Timestamp: 2024-11-26 10:50:16.474 Headers: empty
{"amount":"87274.87","isFlaggedFraud":"0","isFraud":"0","nameDest":"C1746090198","nameOrig":"C899044766","newbalanceDest":"0.0","newbalanceOrig":"9

Offset: 2 Key: empty Timestamp: 2024-11-26 10:50:17.846 Headers: empty
{"amount":"10990.93","isFlaggedFraud":"0","isFraud":"0","nameDest":"M475505653","nameOrig":"C1377243490","newbalanceDest":"0.0","newbalanceOrig":"0

Offset: 3 Key: empty Timestamp: 2024-11-26 10:50:19.858 Headers: empty
```

Figure 2: Kafka Dataset1



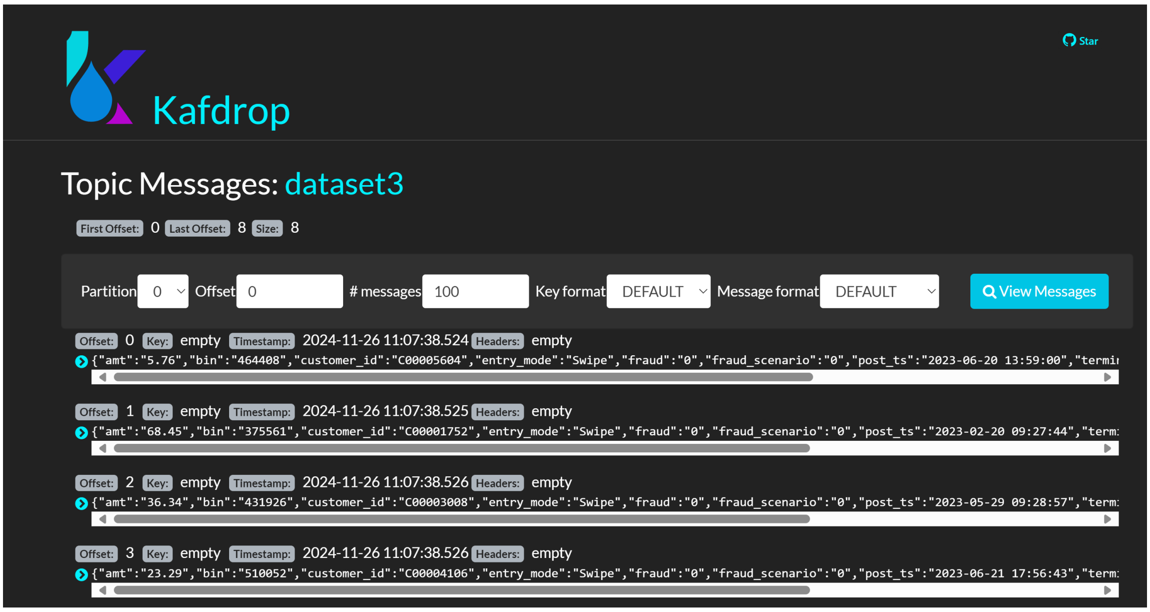
Topic Messages: dataset2

First Offset: 0 Last Offset: 22 Size: 22

Partition 0 Offset 0 # messages 100 Key format DEFAULT Message format DEFAULT

```
Offset: 0 Key: empty Timestamp: 2024-11-26 10:50:15.464 Headers: empty
{"distance_from_home":"0.1767647062","distance_from_last_transaction":"0.1244349809","fraud":"0.0","online_order":"0.0","ratio_to_median_purchase_p

Offset: 1 Key: empty Timestamp: 2024-11-26 10:50:16.474 Headers: empty
{"distance_from_home":"27.3732041476","distance_from_last_transaction":"0.2251119413","fraud":"0.0","online_order":"0.0","ratio_to_median_purchase_

Offset: 2 Key: empty Timestamp: 2024-11-26 10:50:17.856 Headers: empty
{"distance_from_home":"3.7890261686","distance_from_last_transaction":"1.1676261639","fraud":"0.0","online_order":"1.0","ratio_to_median_purchase_p

Offset: 3 Key: empty Timestamp: 2024-11-26 10:50:19.850 Headers: empty
{"distance_from_home":"129.0382748803","distance_from_last_transaction":"25.0816036393","fraud":"1.0","online_order":"1.0","ratio_to_median_purchas
```

Figure 3: Kafka Dataset2



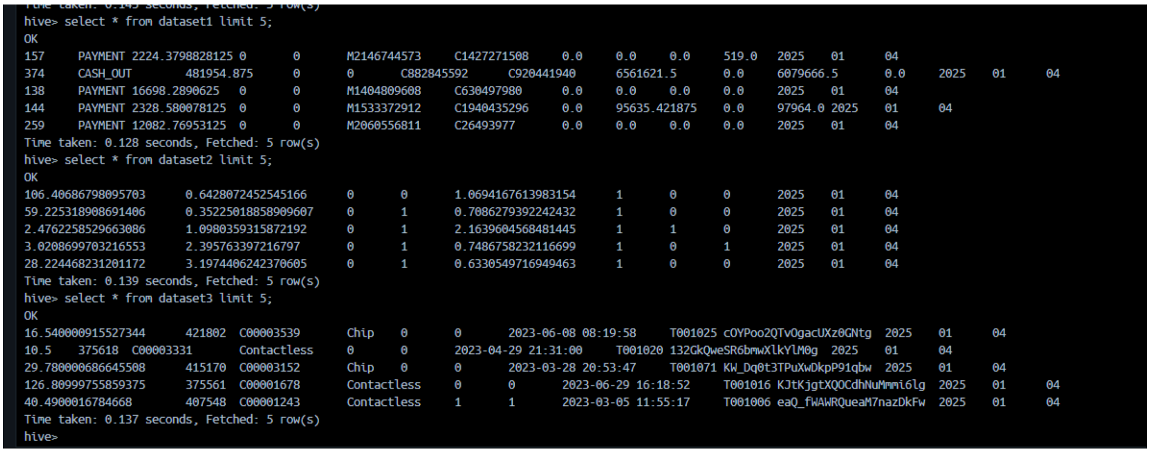
Topic Messages: dataset3

First Offset: 0 Last Offset: 8 Size: 8

Partition 0 Offset 0 # messages 100 Key format DEFAULT Message format DEFAULT

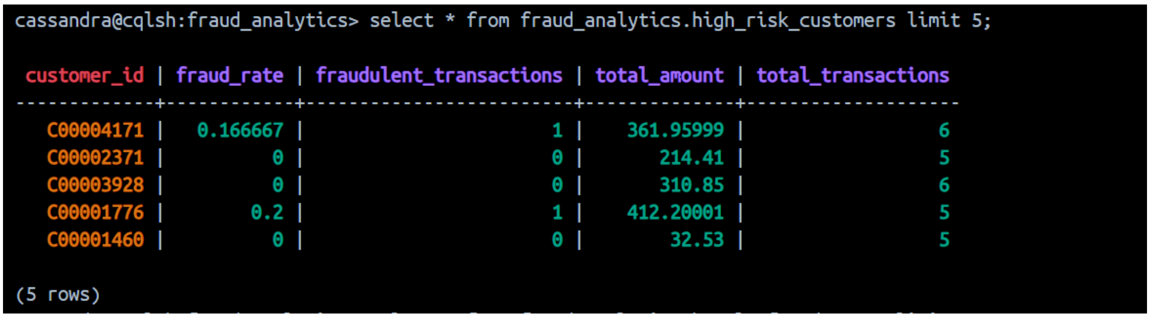
```
Offset: 0 Key: empty Timestamp: 2024-11-26 11:07:38.524 Headers: empty
{"amt":"5.76","bin":"464408","customer_id":"C00005604","entry_mode":"Swipe","fraud":"0","fraud_scenario":"0","post_ts":"2023-06-20 13:59:00","termi

Offset: 1 Key: empty Timestamp: 2024-11-26 11:07:38.525 Headers: empty
{"amt":"68.45","bin":"375561","customer_id":"C00001752","entry_mode":"Swipe","fraud":"0","fraud_scenario":"0","post_ts":"2023-02-20 09:27:44","term

Offset: 2 Key: empty Timestamp: 2024-11-26 11:07:38.526 Headers: empty
{"amt":"36.34","bin":"431926","customer_id":"C00003008","entry_mode":"Swipe","fraud":"0","fraud_scenario":"0","post_ts":"2023-05-29 09:28:57","term

Offset: 3 Key: empty Timestamp: 2024-11-26 11:07:38.526 Headers: empty
{"amt":"23.29","bin":"510052","customer_id":"C00004106","entry_mode":"Swipe","fraud":"0","fraud_scenario":"0","post_ts":"2023-06-21 17:56:43","term
```

Figure 4: Kafka Dataset3



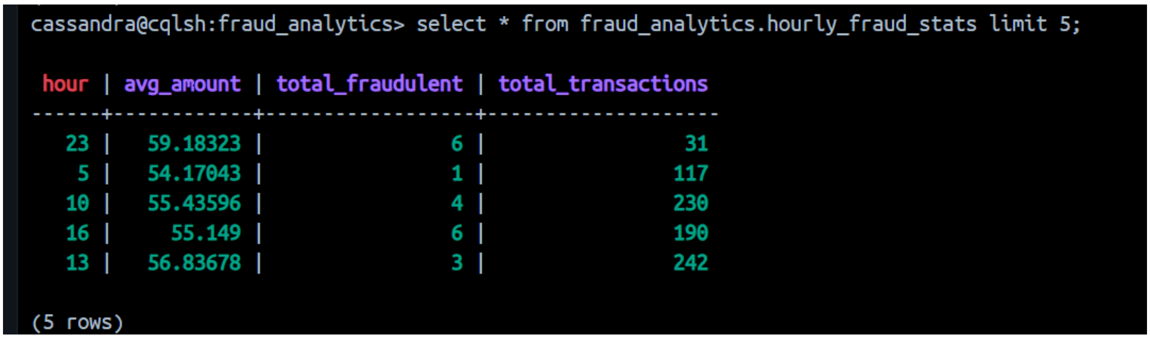
```
Time taken: 0.145 seconds, Fetched: 5 row(s)
hive> select * from dataset1 limit 5;
OK
157     PAYMENT 2224.3798828125 0       0       M2146744573     C1427271508     0.0     0.0     0.0     519.0   2025    01      04
374     CASH_OUT        481954.875      0       0       C882845592      C920441940      6561621.5       0.0     6079666.5       0.0     2025    01      04
138     PAYMENT 16698.2890625   0       0       M1404809608     C630497980      0.0     0.0     0.0     0.0     2025    01      04
144     PAYMENT 2328.580078125  0       0       M1533372912     C1940435296     0.0     95635.421875    0.0     97964.0 2025    01      04
259     PAYMENT 12082.76953125  0       0       M2060556811     C26493977       0.0     0.0     0.0     0.0     2025    01      04
Time taken: 0.128 seconds, Fetched: 5 row(s)
hive> select * from dataset2 limit 5;
OK
106.40686798095703      0.6428072452545166      0       0       1.0694167613983154      1       0       0       2025    01      04
59.225318908691406      0.35225018858909607     0       1       0.7086279392242432      1       0       0       2025    01      04
2.4762258529663086      1.0980359315872192      0       1       2.1639604568481445      1       1       0       2025    01      04
3.0208699703216553      2.395763397216797       0       1       0.7486758232116699      1       0       1       2025    01      04
28.224468231201172      3.1974406242370605      0       1       0.6330549716949463      1       0       0       2025    01      04
Time taken: 0.139 seconds, Fetched: 5 row(s)
hive> select * from dataset3 limit 5;
OK
16.540000915527344      421802  C00003539       Chip    0       0       2023-06-08 08:19:58     T001025 cOYPoo2QTvOgacUXz0GNtg  2025    01      04
10.5    375618  C00003331       Contactless     0       0       2023-04-29 21:31:00     T001020 132GkQweSR6bmwXlkYlM0g  2025    01      04
29.780000686645508      415170  C00003152       Chip    0       0       2023-03-28 20:53:47     T001071 KW_Dq0t3TPuXwDkpP91qbw  2025    01      04
126.80999755859375      375561  C00001678       Contactless     0       0       2023-06-29 16:18:52     T001016 KJtKjgtXQOCdhNuMmmi6lg  2025    01      04
40.4900016784668        407548  C00001243       Contactless     1       1       2023-03-05 11:55:17     T001006 eaQ_fWAWRQueaM7nazDkFw  2025    01      04
Time taken: 0.137 seconds, Fetched: 5 row(s)
hive>
```

Figure 5: Hive data



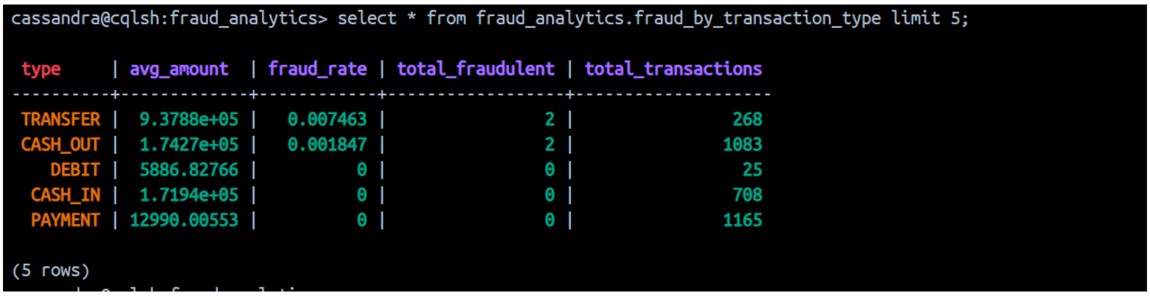
```
cassandra@cqlsh:fraud_analytics> select * from fraud_analytics.high_risk_customers limit 5;

 customer_id | fraud_rate | fraudulent_transactions | total_amount | total_transactions
-------------+------------+-------------------------+--------------+--------------------
   C00004171 |   0.166667 |                       1 |    361.95999 |                  6
   C00002371 |          0 |                       0 |       214.41 |                  5
   C00003928 |          0 |                       0 |       310.85 |                  6
   C00001776 |        0.2 |                       1 |    412.20001 |                  5
   C00001460 |          0 |                       0 |        32.53 |                  5

(5 rows)
```

Figure 6: Cassandra batch processing views



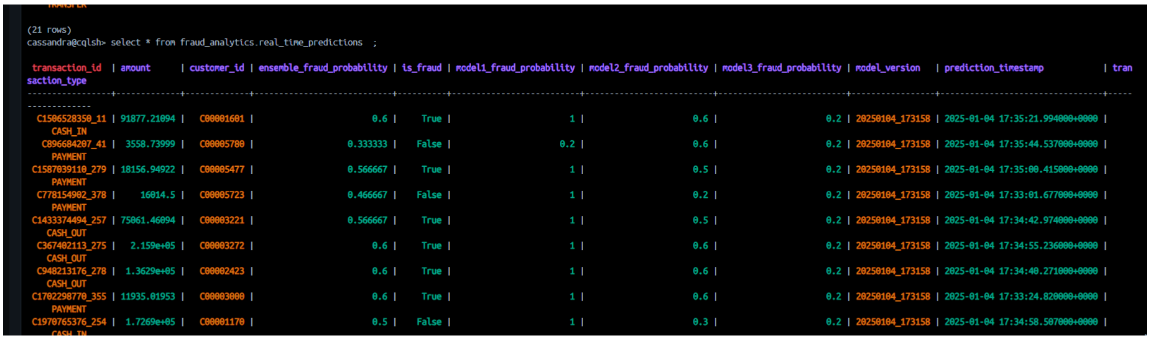
```
cassandra@cqlsh:fraud_analytics> select * from fraud_analytics.hourly_fraud_stats limit 5;

 hour | avg_amount | total_fraudulent | total_transactions
------+------------+------------------+--------------------
   23 |   59.18323 |                6 |                 31
    5 |   54.17043 |                1 |                117
   10 |   55.43596 |                4 |                230
   16 |     55.149 |                6 |                190
   13 |   56.83678 |                3 |                242

(5 rows)
```

Figure 7: Cassandra batch processing views



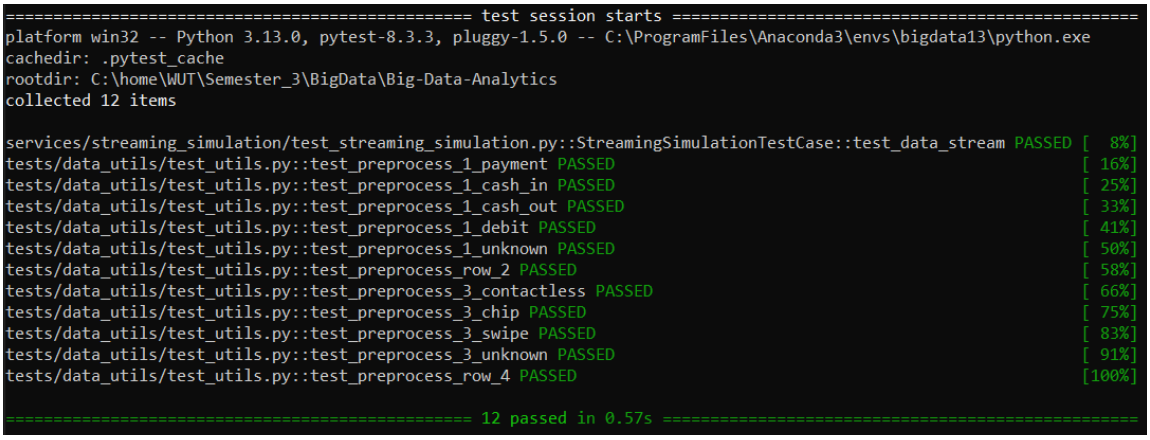
```
cassandra@cqlsh:fraud_analytics> select * from fraud_analytics.fraud_by_transaction_type limit 5;

 type     | avg_amount  | fraud_rate | total_fraudulent | total_transactions
----------+-------------+------------+------------------+--------------------
 TRANSFER |  9.3788e+05 |   0.007463 |                2 |                268
 CASH_OUT |  1.7427e+05 |   0.001847 |                2 |               1083
    DEBIT |  5886.82766 |          0 |                0 |                 25
  CASH_IN |  1.7194e+05 |          0 |                0 |                708
  PAYMENT | 12990.00553 |          0 |                0 |               1165

(5 rows)
```

Figure 8: Cassandra batch processing views



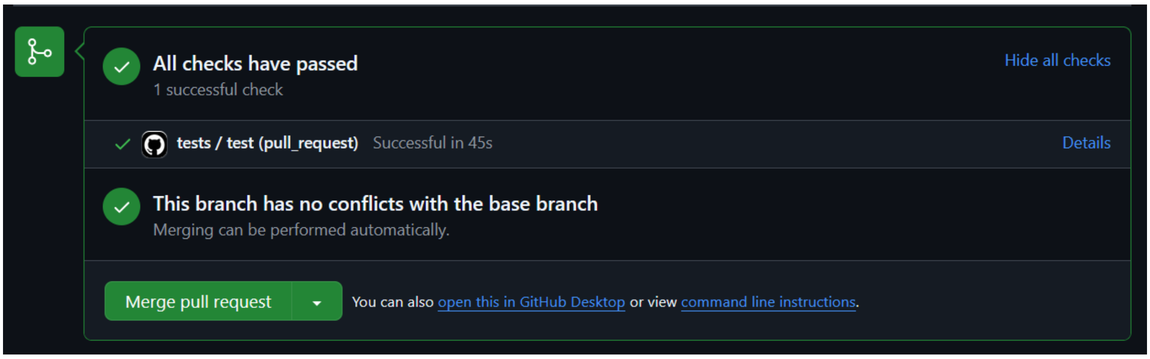
```
      TRANSFER

(21 rows)
cassandra@cqlsh> select * from fraud_analytics.real_time_predictions  ;

 transaction_id  | amount      | customer_id | ensemble_fraud_probability | is_fraud | model1_fraud_probability | model2_fraud_probability | model3_fraud_probability | model_version   | prediction_timestamp            | tran
saction_type
-----------------+-------------+-------------+----------------------------+----------+--------------------------+--------------------------+--------------------------+-----------------+---------------------------------+-----
-------------
  C1506528350_11 | 91877.21094 |   C00001601 |                        0.6 |     True |                        1 |                      0.6 |                      0.2 | 20250104_173158 | 2025-01-04 17:35:21.994000+0000 |
       CASH_IN
   C896684207_41 |  3558.73999 |   C00005780 |                   0.333333 |    False |                      0.2 |                      0.6 |                      0.2 | 20250104_173158 | 2025-01-04 17:35:44.537000+0000 |
       PAYMENT
 C1587039110_279 | 18156.94922 |   C00005477 |                   0.566667 |     True |                        1 |                      0.5 |                      0.2 | 20250104_173158 | 2025-01-04 17:35:00.415000+0000 |
       PAYMENT
  C778154902_378 |     16014.5 |   C00005723 |                   0.466667 |    False |                        1 |                      0.2 |                      0.2 | 20250104_173158 | 2025-01-04 17:33:01.677000+0000 |
       PAYMENT
 C1433374494_257 | 75061.46094 |   C00003221 |                   0.566667 |     True |                        1 |                      0.5 |                      0.2 | 20250104_173158 | 2025-01-04 17:34:42.974000+0000 |
      CASH_OUT
  C367402113_275 |   2.159e+05 |   C00003272 |                        0.6 |     True |                        1 |                      0.6 |                      0.2 | 20250104_173158 | 2025-01-04 17:34:55.236000+0000 |
      CASH_OUT
  C948213176_278 |  1.3629e+05 |   C00002423 |                        0.6 |     True |                        1 |                      0.6 |                      0.2 | 20250104_173158 | 2025-01-04 17:34:40.271000+0000 |
      CASH_OUT
 C1702298770_355 | 11935.01953 |   C00003000 |                        0.6 |     True |                        1 |                      0.6 |                      0.2 | 20250104_173158 | 2025-01-04 17:33:24.820000+0000 |
       PAYMENT
 C1970765376_254 |  1.7269e+05 |   C00001170 |                        0.5 |    False |                        1 |                      0.3 |                      0.2 | 20250104_173158 | 2025-01-04 17:34:58.507000+0000 |
       CASH_IN
```

Figure 9: Cassandra stream processing views



```
================================================ test session starts ================================================
platform win32 -- Python 3.13.0, pytest-8.3.3, pluggy-1.5.0 -- C:\ProgramFiles\Anaconda3\envs\bigdata13\python.exe
cachedir: .pytest_cache
rootdir: C:\home\WUT\Semester_3\BigData\Big-Data-Analytics
collected 12 items

services/streaming_simulation/test_streaming_simulation.py::StreamingSimulationTestCase::test_data_stream PASSED [  8%]
tests/data_utils/test_utils.py::test_preprocess_1_payment PASSED                                                 [ 16%]
tests/data_utils/test_utils.py::test_preprocess_1_cash_in PASSED                                                 [ 25%]
tests/data_utils/test_utils.py::test_preprocess_1_cash_out PASSED                                                [ 33%]
tests/data_utils/test_utils.py::test_preprocess_1_debit PASSED                                                   [ 41%]
tests/data_utils/test_utils.py::test_preprocess_1_unknown PASSED                                                 [ 50%]
tests/data_utils/test_utils.py::test_preprocess_row_2 PASSED                                                     [ 58%]
tests/data_utils/test_utils.py::test_preprocess_3_contactless PASSED                                             [ 66%]
tests/data_utils/test_utils.py::test_preprocess_3_chip PASSED                                                    [ 75%]
tests/data_utils/test_utils.py::test_preprocess_3_swipe PASSED                                                   [ 83%]
tests/data_utils/test_utils.py::test_preprocess_3_unknown PASSED                                                 [ 91%]
tests/data_utils/test_utils.py::test_preprocess_row_4 PASSED                                                     [100%]

================================================ 12 passed in 0.57s ================================================
```

Figure 10: Unit testing result



All checks have passed

1 successful check

✓ tests / test (pull_request) Successful in 45s

This branch has no conflicts with the base branch

Merging can be performed automatically.

You can also open this in GitHub Desktop or view command line instructions.

Figure 11: GitHub checks before merge