Case Study: Synthetic Financial

Domain: Financial

There is a lack of publicly available datasets on financial services and especially in the emerging mobile money transactions domain. Financial datasets are important for many researchers at performing research in the domain of fraud detection.

Tasks: Now with data pipeline ready, you are required to develop the model and predict the fraud using spark streaming.

Module:Mod11CS2_FraudModel-Final

1. Explore the dataset and develop a model to predict the fraud.
2. Develop the application to train the model and persist the model to disk.

Module: Mod11CS2_Consumer

3. Create a new spark streaming application for the same

Module : fraud.conf

4. Application will connect to the flume to retrieve the data

Module: Mod11CS2_Consumer

5. Load the model
6. Predict the fraud and print the result to the logs
7. Test the application by sending dummy data rows from the consumer

Data Passed via Flume

```
^CConnection closed by foreign host.
[edureka_524533@ip-20-0-31-82 ~]$ telnet localhost 44444
Trying ::1...
telnet: connect to address ::1: Connection refused
Trying 127.0.0.1...
Connected to localhost.
Escape character is '^]'.
1,PAYMENT,9839.64,C1231006815,170136.0,160296.36,M1979787155,0.0,0.0,0,0
\OK
1,PAYMENT,9839.64,C1231006815,170136.0,160296.36,M1979787155,0.0,0.0,0,0
OK
1,PAYMENT,9839.64,C1231006815,170136.0,160296.36,M1979787155,0.0,0.0,0,0
OK
1,CASH_OUT,181.0,C840083671,181.0,0.0,C38997010,21182.0,0.0,1,0
OK
1,TRANSFER,181.0,C1305486145,181.0,0.0,C553264065,0.0,0.0,1,0
OK
1,TRANSFER,181.0,C1305486145,181.0,0.0,C553264065,0.0,0.0,1,0
OK
1,CASH_OUT,181.0,C840083671,181.0,0.0,C38997010,21182.0,0.0,1,0
OK
1,PAYMENT,9839.64,C1231006815,170136.0,160296.36,M1979787155,0.0,0.0,0,0
OK
1,CASH_OUT,181.0,C840083671,181.0,0.0,C38997010,21182.0,0.0,1,0
OK
1,CASH_OUT,181.0,C840083671,181.0,0.0,C38997010,21182.0,0.0,1,0
OK
1,TRANSFER,215310.3,C1670993182,705.0,0.0,C1100439041,22425.0,0.0,0,0
OK
```

Written to HDFS

Home / user / edureka_524533 / Flume_Fraud / 2019-07-30 / **events-.1564503820981.csv**

```
1,CASH_OUT,181.0,C840083671,181.0,0.0,C38997010,21182.0,0.0,1,0
```

Processed by Pyspark Application:

```
ssc.awaitTermination()  # Wait for the computation to terminate
```

```
-------------------------------------------
Time: 2019-07-30 17:02:00
-------------------------------------------
1,CASH_OUT,181.0,C840083671,181.0,0.0,C38997010,21182.0,0.0,1,0
1,TRANSFER,215310.3,C1670993182,705.0,0.0,C1100439041,22425.0,0.0,0,0

=== RDD Found ===
root
 |-- line: string (nullable = true)

root
 |-- line: string (nullable = true)
 |-- step: integer (nullable = true)
 |-- type: string (nullable = true)
 |-- amount: double (nullable = true)
 |-- nameOrig: string (nullable = true)
 |-- oldbalanceOrg: double (nullable = true)
 |-- newbalanceOrig: double (nullable = true)
 |-- nameDest: string (nullable = true)
 |-- oldbalanceDest: double (nullable = true)
 |-- newbalanceDest: double (nullable = true)
 |-- isFraud: integer (nullable = true)
 |-- isFlaggedFraud: integer (nullable = true)

+-------+----------+
|isFraud|prediction|
+-------+----------+
|      1|       0.0|
|      0|       1.0|
+-------+----------+

-------------------------------------------
Time: 2019-07-30 17:04:00
-------------------------------------------

==== EMPTY ====
```

Data :

Paysim synthetic dataset of mobile money transactions. Each step represents an hour of simulation. This dataset is scaled down 1/4 of the original data-set which is presented in the paper "Paysim: A financial mobile money simulator for fraud detection”.