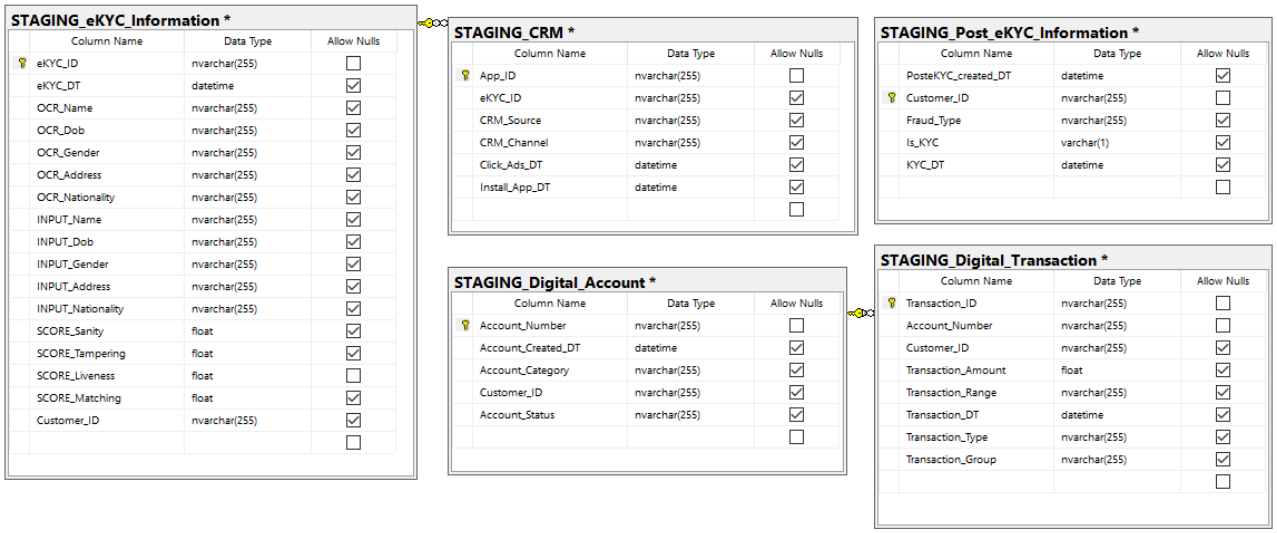
Quản lý Metadata

1. Các bên liên quan:

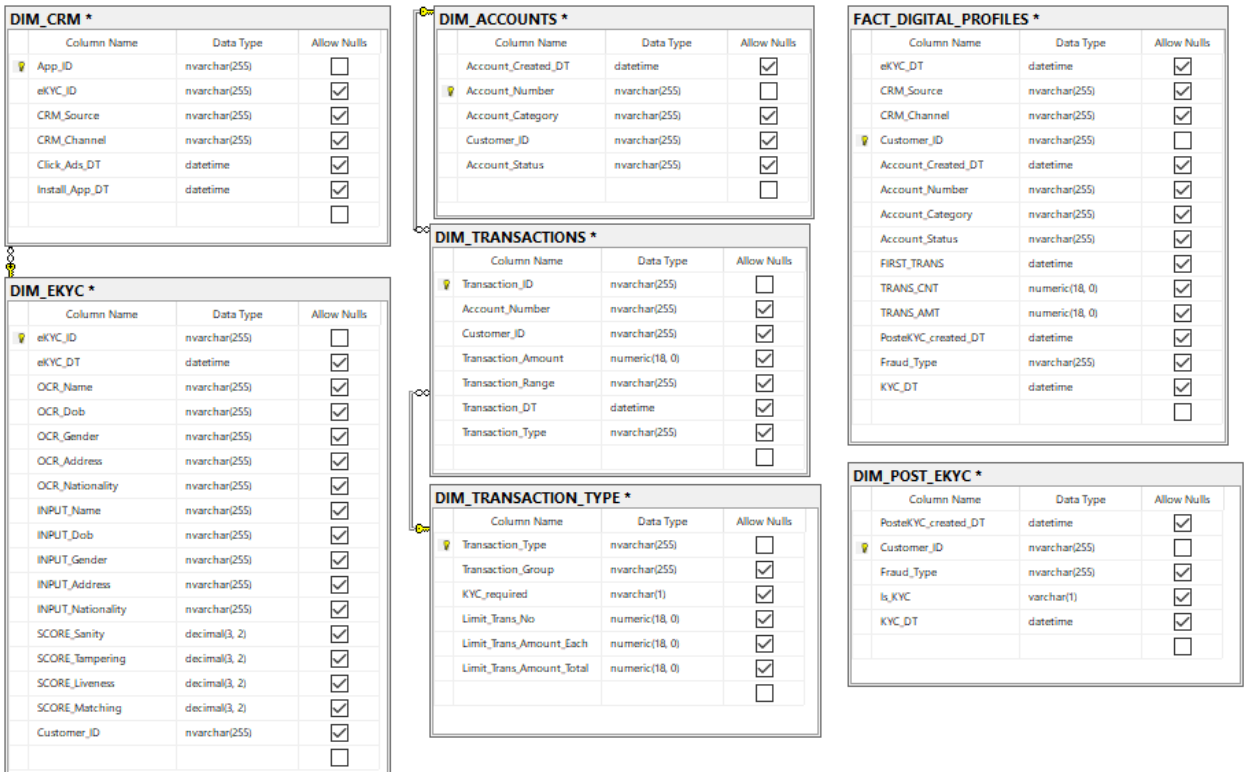
Vị trí	Vai trò
Business Analyst	Cung cấp các yêu cầu về xây dựng hệ thống Data warehouse.
Data Architect	Thiết kế, xây dựng hệ thống Data warehouse.
Data Engineer	Tiền xử lý và tích hợp dữ liệu vào Data warehouse.
Data Govenance Analyst	Quản lý chất lượng dữ liệu.

2. Tạo Schema cho tầng STAGING:



Hình 2-1: STAGING Schema.

3. Tạo schema cho tầng RECONCILIATION:



Hình 3-1: RECILINATION Schema.

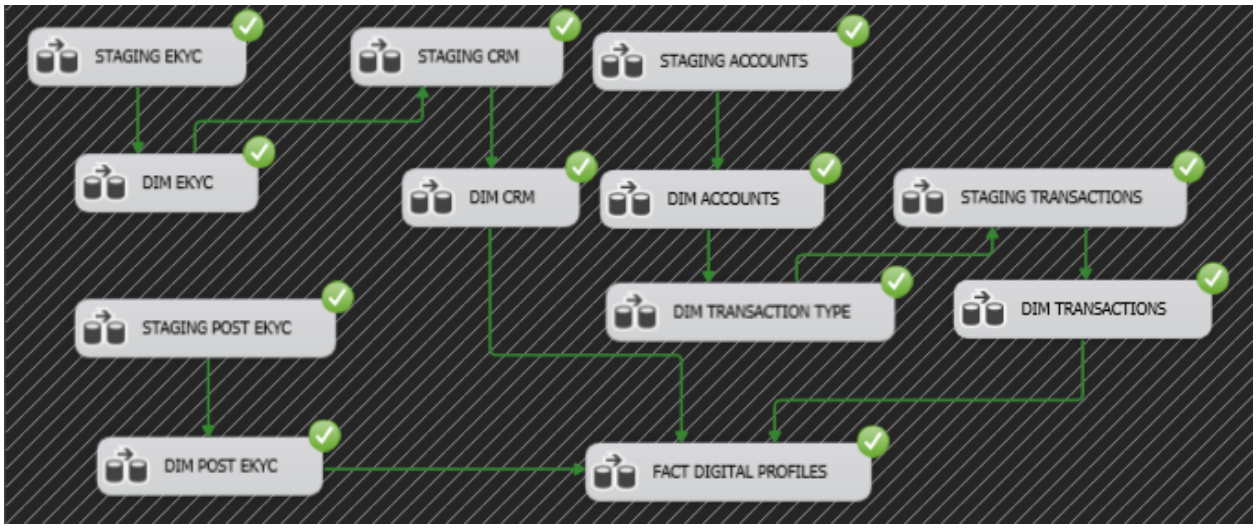
4. Tích hợp dữ liệu bằng ETL Tools:

4.1. Chuyển đổi và lọc dữ liệu:

Có thể tham khảo file Excel sau để hiểu rõ hơn về các yêu cầu của BA về mapping dữ liệu giữa các bảng:

Source-Target_Mapping.xlsx

4.2. Tiến hành ETL data:



Hình 4-1: Kết quả thực hiện.

Ngoài các bảng mapping 1-1 ra thì có các bảng sau cần phải lọc theo yêu cầu:

- [dbo].[STAGING_eKYC_Information]:

```
SELECT [eKYC_ID]
      ,[eKYC_DT]
      ,JSON_VALUE([OCR_INFO],'$.name') [OCR_Name]
         ,JSON_VALUE([OCR_INFO],'$.dob') [OCR_Dob]
         ,JSON_VALUE([OCR_INFO],'$.gender') [OCR_Gender]
         ,JSON_VALUE([OCR_INFO],'$.nationality') [OCR_Nationality]
         ,JSON_VALUE([OCR_INFO],'$.address') [OCR_Address]
         ,JSON_VALUE([INPUT_INFO],'$.name') [INPUT_Name]
         ,JSON_VALUE([INPUT_INFO],'$.dob') [INPUT_Dob]
         ,JSON_VALUE([INPUT_INFO],'$.gender') [INPUT_Gender]
         ,JSON_VALUE([INPUT_INFO],'$.address') [INPUT_Address]
         ,JSON_VALUE([INPUT_INFO],'$.nationality') [INPUT_Nationality]
      ,CAST([SANITY_SCORE] AS decimal(3,2)) [SCORE_Sanity]
         ,CAST([SANITY_SCORE] AS decimal(3,2)) [SCORE_Tampering]
         ,CAST([SANITY_SCORE] AS decimal(3,2)) [SCORE_Liveness]
         ,CAST([SANITY_SCORE] AS decimal(3,2)) [SCORE_Matching]
      ,[CUSTOMER_ID]
   FROM [ONBOARDING].[dbo].[ONBOARDING_Data]
```

- [dbo].[STAGING_Digital_Account]:

```
SELECT  [CREATED_DT]
      ,[Transaction_Account]
      ,[Account_Category]
      ,[CUSTOMER_ID]
      ,[ACCOUNT_STATUS]
FROM [CORE_T24].[dbo].[T24_ACCOUNT]
   WHERE Account_Category in ('1001','1002')
```

- [dbo].[DIM_TRANSACTION_TYPE]:

```
SELECT  [Transaction_Type]
      ,[Transaction_Group]
FROM [CORE_T24].[dbo].[T24_TRANSACTION]
GROUP BY  [Transaction_Type]
        ,[Transaction_Group]
```

- [dbo].[STAGING_Digital_Account]

```
SELECT [Transaction_ID]
      ,[Transaction_Account]
      ,[CUSTOMER_ID]
      --,[Channel]
         ,[Transaction_Amount]
         , IIF([Transaction_Amount] < 1000000, 'LOW'
                     , IIF([Transaction_Amount] < 10000000, 'MEDIUM LOW'
                           , IIF([Transaction_Amount] < 100000000, 'MEDIUM HIGH',
'HIGH')))   [Transaction_Range]
         ,[Transaction_DT]
      ,[Transaction_Type]
      ,[Transaction_Group]
  FROM [CORE_T24].[dbo].[T24_TRANSACTION]
     WHERE [Channel] = 'APP'
```

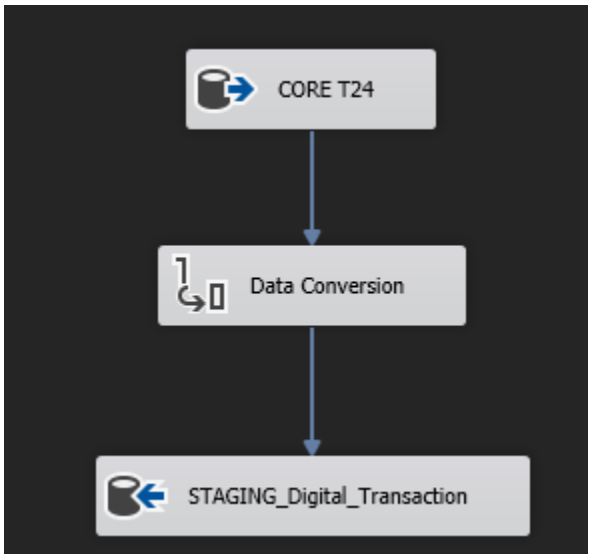
- FACT

```
SELECT e.[eKYC_DT], e.[Customer_ID]
                     , c.[CRM_Source], c.[CRM_Channel]
                     , a.[Account_Created_DT], a.[Account_Number],
a.[Account_Category], a.[Account_Status]
                     , t.FIRST_TRANS, t.TRANS_CNT, t.TRANS_AMT
                     , p.[PosteKYC_created_DT]
                     , p.[Fraud_Type]
                     , p.[KYC_DT]
                     --,
       FROM [DWH].[dbo].[DIM_EKYC] e
       LEFT JOIN [DWH].[dbo].[DIM_CRM] c on e.[eKYC_ID] = c.[eKYC_ID]
       LEFT JOIN [DWH].[dbo].[DIM_ACCOUNTS] a on e.[Customer_ID] = a.[Customer_ID]
       LEFT JOIN (SELECT [Customer_ID],[Account_Number]
                                       , min([Transaction_DT]) FIRST_TRANS
                                       , count(distinct [Transaction_ID])
TRANS_CNT
                                       , sum([Transaction_Amount]) TRANS_AMT
                            FROM [DWH].[dbo].[DIM_TRANSACTIONS]
                            GROUP BY  [Customer_ID],[Account_Number]
                                  ) t on e.[Customer_ID] = t.[Customer_ID]
                                  AND a.[Account_Number] = t.[Account_Number]
       LEFT JOIN [DWH].[dbo].[DIM_POST_EKYC] p on   e.[Customer_ID] =
p.[Customer_ID]
       WHERE e.[Customer_ID] IS NOT NULL
```

4.3. Một vài vấn đề gặp phải trong quá trình ETL dữ liệu:

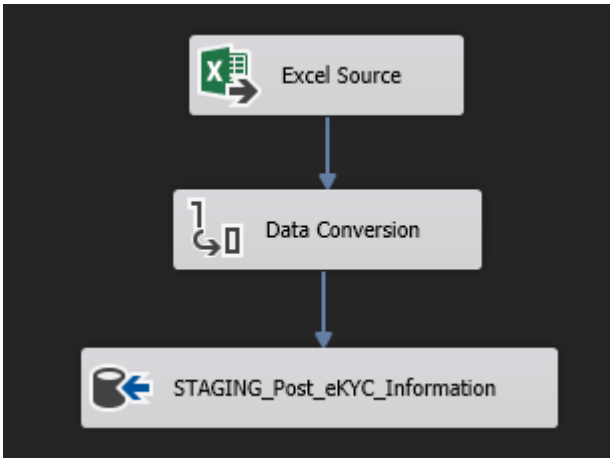
- Xung đột về kiểu dữ liệu:

- Trong quá trình tải dữ liệu từ CORE_T24 vào bảng [STAGING Digital Transaction]. Vì dữ liệu đầu vào (CORE_T24) ở cột Transaction_Range là kiểu String, còn cột Transaction_Range của bảng STAGING Digital Transaction trong database là nvarchar vì vậy ta cần phải chuyển dữ liệu đầu vào sang Unicode string [DT_WSTR].



Hình 4-2:STAGING_Digital_Transaction Data Flow

- Trong quá trình tải dữ liệu từ nguồn là một file Excel vào bảng [STAGING_Post_eKYC_Information]. Trong trường hợp này, dữ liệu đầu vào của cột IS_KYC là dạng String, nhưng trong database, thì kiểu dữ liệu của nó là varchar, tức là ASCII vì vậy ta cần phải đổi code page của phần Data Convertion từ 1258 sang 1252.



Hình 4-3: STAGING_Post_eKYC_Information Data Flow

5. Đánh giá chất lượng dữ liệu:

5.1. Báo cáo về độ chính xác của dữ liệu:

- Viết một câu query với các trường sau:
  - IS_TIME_ALERT: xem xét các mốc thời gian mà khách hàng thực hiện các thao tác có hợp lý hay không (click -> install -> eKYC -> Account Created -> First Transaction)
  - IS_SCORE_ALERT: khách hàng đã đạt KYC nhưng các điểm số của khách hàng lại không đạt.
  - IS_ONFO_ALERT: khách hàng đạt KYC nhưng dữ liệu từ OCR không đồng nhất với dữ liệu mà khách hàng nhập vào.
  - IS_CATEGORY_ALERT: tài khoản của khách hàng không phải loại 1001 và 1002
  - Fraud_Type: phân loại khách hàng

```
select F.Customer_ID
       , IIF(C.Click_Ads_DT>C.Install_App_DT OR C.Click_Ads_DT > F.eKYC_DT OR
C.Click_Ads_DT > F.Account_Created_DT OR C.Click_Ads_DT > F.FIRST_TRANS
              OR C.Install_App_DT> F.eKYC_DT OR C.Install_App_DT > F.Account_Created_DT
OR C.Install_App_DT > F.FIRST_TRANS
              OR F.eKYC_DT > F.Account_Created_DT OR F.eKYC_DT > F.FIRST_TRANS
              OR F.Account_Created_DT > F.FIRST_TRANS, 1,0) IS_TIME_ALERT
       , IIF(E.[SCORE_Sanity] < 0.85
              OR E.[SCORE_Tampering] < 0.85
              OR E.[SCORE_Liveness] < 0.85
              OR E.[SCORE_Matching] < 0.85 , 1 , 0) IS_SCORE_ALERT
       , IIF( [OCR_Name]<>[INPUT_Name]
              OR [OCR_Dob] <> [INPUT_Dob]
              OR [OCR_Gender] <> [INPUT_Gender]
        OR [OCR_Address] <> [INPUT_Address]
        OR [OCR_Nationality]<>[INPUT_Nationality],1,0) IS_INFO_ALERT
       , IIF(F.Account_Category NOT IN ('1001','1002'),1,0) IS_CATEGORY_ALERT
       , F.Fraud_Type
FROM FACT_DIGITAL_PROFILES F
LEFT JOIN DIM_EKYC E ON E.Customer_ID=F.Customer_ID
LEFT JOIN DIM_CRM C ON C.eKYC_ID=E.eKYC_ID
```

Với dữ liệu có được từ câu query trên và dùng Excel Piviot Table ta có 2 bảng sau:

Row Labels	Sum of IS_TIME_ALERT	Sum of IS_SCORE_ALERT	Sum of IS_INFO_ALERT	Sum of IS_CATEGORY_ALERT	Count of Customer_ID
CHECKED	116	157	0	0	3553
FRAUD	138	271	0	0	6000
RISK	43	75	0	0	1783
NORMAL	551	1504	0	0	35780
Grand Total	848	2007	0	0	47116

Hình 5-1: Báo cáo độ chính xác của dữ liệu.

Row Labels	Sum of IS_TIME_ALER	Sum of IS_SCORE_ALERT	Sum of IS_INFO_ALER	Sum of IS_CATEGORY_ALER	Count of Customer_I
CHECKED	3%	4%	0	0	3553
FRAUD	2%	5%	0	0	6000
RISK	2%	4%	0	0	1783
NORMAL	2%	4%	0	0	35780

Hình 5-2: Báo cáo độ chính xác của dữ liệu (phần trăm).

- Đánh giá:
  - Qua báo cáo trên ta có thể thấy các trường IS_INFO_ALERT và IS_CATEGORY_ALERT không có khách hàng nào vi phạm cả, nó thể hiện sự nhất quán giữa dữ liệu OCR và dữ liệu mà khách hàng nhập vào. Ngoài ra, tất cả các khách hàng đều có tài khoản thuộc loại 1001 và 1002.
  - Tuy nhiên đối với trường IS_SCORE_ALERT và IS_TIME_ALERT thì số lượng khách hàng vi phạm khá cao, đặc biệt đối với khách hàng NORMAL ở trường IS_SCORE_ALERT lên đến 1504 khách hàng, vì vậy cần phải tiến hành rà soát và kiểm tra lại tập khách hàng này để tránh việc họ mang lại giá trị không tốt cho ngân hàng. Ngoài ra khi tính toán theo phần trăm khách hàng so với tổng số thì trường IS_SCORE_ALERT cũng chiếm tỉ lệ khá cao (4-5%) vì vậy cần phải kiểm tra lại hệ thống OCR có hoạt động tốt không hay tiêu chuẩn về các điểm số chưa phù hợp.

5.2. Báo cáo về tính nhất quán của dữ liệu:

- Làm bảng báo cáo với các cột sau:
  - eKYC_MONTH: tháng thực hiện eKYC
  - FRAUD_NOT_CLOSED: khách hàng là Fraud nhưng chưa đóng tài khoản.
  - RISK_NOT_SUSPENDED: khách hàng là Risk nhưng chưa đình chỉ hoạt động hoặc cảnh báo.
  - CHECK_NOT_ACTIVE: khách hàng bình thường nhưng tài khoảng không được cho phép hoạt động.
  - RISK_TRANS: những giao dịch thực hiện bởi khách hàng là RISK đã thực hiện những giao dịch đáng ra đã bị giới hạn.
  - FRAUD_TRANS: những giao dịch thực hiện bởi khách hàng là FRAUD.

```
SELECT CONVERT(VARCHAR(6),EKYC_DT,112) eKYC_MONTH
              , F.Customer_ID
              , IIF(Fraud_Type = 'FRAUD' AND Account_Status <> 'CLOSED',1,0)
FRAUD_NOT_CLOSED
              , IIF(Fraud_Type = 'RISK' AND Account_Status <> 'SUSPENDED',1,0)
RISK_NOT_SUSPENDED
              , IIF(Fraud_Type = 'CHECKED' AND Account_Status <> 'ACTIVE',1,0)
CHECK_NOT_ACTIVE
              , IIF(FRAUD_TRANS.Account_Number IS NOT NULL,1,0) FRAUD_TRANS
              , FRAUD_TRANS_ID
              , FRAUD_TRANS_AMOUNT
```

```
                , FRAUD_TRANS_GROUP
                , FRAUD_TRANS_RANGE
                , IIF(RISK_TRANS.Account_Number IS NOT NULL,1,0) RISK_TRANS
                , RISK_TRANS_ID
                , RISK_TRANS_AMOUNT
                , RISK_TRANS_GROUP
                , RISK_TRANS_RANGE
        FROM FACT_DIGITAL_PROFILES F
        LEFT JOIN (
                SELECT F.Customer_ID
                        , T.Transaction_ID FRAUD_TRANS_ID
                        , T.Account_Number
                        , T.Transaction_Amount FRAUD_TRANS_AMOUNT
                        , TT.Transaction_Group FRAUD_TRANS_GROUP
                        , T.Transaction_Range FRAUD_TRANS_RANGE
                FROM DIM_TRANSACTIONS T LEFT JOIN FACT_DIGITAL_PROFILES F ON
F.Account_Number=T.Account_Number
                LEFT JOIN DIM_TRANSACTION_TYPE TT ON TT.Transaction_Type=T.Transaction_Type
                WHERE T.Transaction_DT>=F.PosteKYC_created_DT
                        AND F.Fraud_Type='FRAUD'
        ) FRAUD_TRANS ON F.Account_Number=FRAUD_TRANS.Account_Number
        LEFT JOIN (
                SELECT F.Customer_ID
                        , T.Transaction_ID RISK_TRANS_ID
                        , T.Account_Number
                        , T.Transaction_Type RISK_TRANS_AMOUNT
                        , TT.Transaction_Group RISK_TRANS_GROUP
                        , T.Transaction_Range RISK_TRANS_RANGE
                FROM DIM_TRANSACTIONS T LEFT JOIN FACT_DIGITAL_PROFILES F ON
F.Account_Number=T.Account_Number
                LEFT JOIN DIM_TRANSACTION_TYPE TT ON TT.Transaction_Type=T.Transaction_Type
                WHERE T.Transaction_DT>=F.PosteKYC_created_DT
                        AND TT.Transaction_Group='DEPOSIT'
                        AND F.Fraud_Type='RISK'
        ) RISK_TRANS ON F.Account_Number=RISK_TRANS.Account_Number
```

Với dữ liệu có được từ câu query trên và dùng Excel Piviot Table ta có báo cáo sau:

eKYC_MONTH	FRAUD_NOT_CLOSED	RISK_NOT_SUSPENDED	CHECK_NOT_ACTIVE	FRAUD_TRANS	RISK_TRANS
202201	0	0	0	0	0
202202	0	0	0	1	0
202203	0	0	0	3	0
202204	0	0	0	8	1
202205	0	0	0	17	0
202206	0	0	0	29	1
202207	0	0	0	43	2
202208	0	0	0	61	7
202209	0	0	0	128	5
202210	0	0	0	186	16
202211	0	0	0	309	20
202212	0	0	0	549	33
Grand Total	0	0	0	1334	85

Hình 5-3: Báo cáo số lượng các giao dịch của khách hàng là FRAUD và RISK

eKYC_MONTH	DEPOSIT				PAYMENT						TRANSFER							
	MEDIUM HIGH		HIGH		MEDIUM LOW		MEDIUM HIGH		HIGH		LOW		MEDIUM LOW		MEDIUM HIGH		HIGH	
	CNT	Total_AMT	CNT	Total_AMT	CNT	Total_AMT	CNT	Total_AMT	CNT	Total_AMT	CNT	Total_AMT	CNT	Total_AMT	CNT	Total_AMT	CNT	Total_AMT
202202															1	61.776.328		
202203									2	815.158.406					1	75.861.090		
202204	1	79.877.054	2	882.304.902					4	1.016.052.012					1	83.392.978		
202205			2	781.683.503					9	2.588.012.330					2	135.887.112	4	544.450.362
202206	1	68.705.325	5	2.049.743.148	1	6.352.865	1	51.174.150	7	2.236.631.720	1	286.012	2	5.250.112	7	312.659.833	4	535.991.359
202207			9	3.641.863.800			6	271.192.674	20	6.706.762.574					3	106.261.696	5	807.861.491
202208	1	24.425.022	7	4.587.574.226	1	4.574.979	3	167.995.998	14	4.285.078.231			1	8.496.865	14	607.623.916	20	3.037.604.672
202209	4	197.301.506	32	15.301.070.084	1	6.634.659	7	322.158.384	42	13.539.664.874			6	23.892.958	15	739.330.351	21	3.217.045.089
202210	4	185.898.646	28	14.013.686.302	1	1.547.964	8	538.629.814	82	24.746.129.500			2	13.998.147	24	1.460.453.600	37	5.795.739.942
202211	4	245.603.539	51	30.638.784.816	3	22.711.394	19	855.421.289	113	32.540.314.548			3	13.102.249	58	3.167.823.602	58	8.764.453.123
202212	10	519.229.480	114	68.330.383.881	3	16.183.861	45	2.480.221.259	183	57.623.433.939			11	70.804.703	97	5.550.954.519	86	12.922.210.885
Grand Total	25	1.321.040.572	250	140.227.094.662	10	58.005.722	89	4.686.793.568	476	146.097.238.134	1	286.012	25	135.545.034	223	12.302.025.025	235	35.625.356.923

Hình 5-4: Báo cáo chi tiết về các loại giao dịch của các khách hàng là FRAUD