2. Documento de Testes (unitário e integrado)

O processo de teste serve para validar o processo de carga de dados. Serão realizados teste unitário e teste integrado de cada uma das cargas.

Teste unitário

O teste unitário valida se todas as linhas do arquivo foram inseridos no Data Lake.

Teste integrado

O teste integrado valida o a integridade de relacionamento entre os dados.

Importação de bibliotecas

In [1]:
```
import os
import shutil
from pyspark.sql import SparkSession
from pyspark.sql.functions import from_unixtime, col, to_timestamp, coalesce
from pyspark.sql.types import StringType, IntegerType, LongType, DecimalType, DateType
```

Variaveis do projeto

In [2]:
```
#Diretorio dos arquivos csv
v_diretorio_csv='/usr/local/spark/csv/'

#Variaveis de conexao com postgres
v_caminho_jar_postgres='/home/jovyan/work/jars/postgresql-9.4.1207.jar'
v_url_jdbc='jdbc:postgresql://postgres/projeto'
v_user_jdbc='airflow'
v_pass_jdbc='airflow'
```

Criando sessao e contexto

In [3]:
```
spark = (SparkSession
         .builder
         .master('local')
         .appName('load-postgres')
         # Add postgres jar
         .config('spark.driver.extraClassPath', v_caminho_jar_postgres)
         .getOrCreate())
sc = spark.sparkContext
```

```
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/usr/local/spark-
3.1.2-bin-hadoop3.2/jars/spark-unsafe_2.12-3.1.2.jar) to constructor java.nio.DirectByteBuffer
(long,int)
WARNING: Please consider reporting this to the maintainers of org.apache.spark.unsafe.Platform
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access oper
ations
WARNING: All illegal access operations will be denied in a future release
22/12/22 20:31:44 WARN NativeCodeLoader: Unable to load native-hadoop library for your platfor
m... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
22/12/22 20:31:46 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 40
41.
```

Lendo arquivo csv, criando dataframe spark, formatando e criando views

Essa fase do processo, carrega os dados dos arquivos csv em dataframes, formata os campos e cria views para posteriormente serem utilizados na fase de tratamento dos dados.

In [4]:

```
#Dataframe Associado
df_associado_csv = (
    spark.read
    .format('csv')
    .option('header', True)
    .option('delimiter', ';')
    .load(v_diretorio_csv + 'associado.csv')
)

#Definindo o tipo da coluna
df_associado_csv_fmt = (
    df_associado_csv
    .withColumn('id', col('id').cast(IntegerType()))
    .withColumn('idade', col('idade').cast(IntegerType()))
)

#Criando view do dataframe
df_associado_csv_fmt.createOrReplaceTempView('associado_csv')
```

In [5]:

```
df_associado_csv_fmt.show()
```

```
+---+---------+---------+-----+--------------------+
| id|     nome|sobrenome|idade|               email|
+---+---------+---------+-----+--------------------+
|  1|   Alícia|  Cardoso|   29|alícia.cardoso@ho...|
|  2|  Mirella|    Moura|   25|mirella.moura@gma...|
|  3|  Rodrigo|Fernandes|   54|rodrigo.fernandes...|
|  4|   Rebeca|  Cardoso|   59|rebeca.cardoso@te...|
|  5|     Raul|   Barros|   51|raul.barros@yahoo...|
|  6|    Julia|    Nunes|   38|julia.nunes@yahoo...|
|  7|     João|   Miguel|   45|joão.miguel@uol.c...|
|  8|Francisco|    Gomes|   27|francisco.gomes@h...|
|  9| Vinicius|     Lima|   58|vinicius.lima@hot...|
| 10|  Cecília|    Souza|   40|cecília.souza@uol...|
| 11|      Ana|    Julia|   57|ana.julia@yahoo.c...|
| 12|  Anthony|    Neves|   40|anthony.neves@yah...|
| 13|    Lucas|    Costa|   34|lucas.costa@hotma...|
| 14|      Ana| Teixeira|   66|ana.teixeira@hotm...|
| 15|     João|    Lucas|   70|joão.lucas@uol.co...|
| 16|    Bruna|      Luz|   69|bruna.luz@hotmail...|
| 17|    Vitor|     Hugo|   67|vitor.hugo@hotmai...|
| 18|    Sarah|Fernandes|   39|sarah.fernandes@y...|
| 19|  Cecília|Rodrigues|   75|cecília.rodrigues...|
| 20|   Nathan|     Mota|   42|nathan.mota@yahoo...|
+---+---------+---------+-----+--------------------+
only showing top 20 rows
```

In [6]:

```
#Dataframe Conta
df_conta_csv = (
    spark.read
    .format('csv')
    .option('header', True)
    .option('delimiter', ';')
    .load(v_diretorio_csv + 'conta.csv')
)

#Definindo o tipo da coluna
df_conta_csv_fmt = (
    df_conta_csv
    .withColumn('id', col('id').cast(IntegerType()))
    .withColumn('data_criacao', col('data_criacao').cast(DateType()))
    .withColumn('id_associado', col('id_associado').cast(IntegerType()))
)
```

```
#Criando view do dataframe
df_conta_csv_fmt.createOrReplaceTempView('conta_csv')
```

In [7]:

```
df_conta_csv_fmt.show()
```

```
+---+--------------+-----------+------------+
| id|          tipo|data_criacao|id_associado|
+---+--------------+-----------+------------+
|  1|Conta Corrente|  2019-03-28|           1|
|  2|Conta Corrente|  2021-04-02|           2|
|  3|Conta Corrente|  2019-05-24|           3|
|  4|Conta Corrente|  2018-10-22|           4|
|  5|Conta Corrente|  2022-11-29|           5|
|  6|Conta Corrente|  2018-05-26|           6|
|  7|Conta Corrente|  2020-08-23|           7|
|  8|Conta Corrente|  2019-02-16|           8|
|  9|Conta Corrente|  2021-03-09|           9|
| 10|Conta Corrente|  2022-04-09|          10|
| 11|Conta Corrente|  2019-10-08|          11|
| 12|Conta Corrente|  2022-04-28|          12|
| 13|Conta Corrente|  2019-02-15|          13|
| 14|Conta Corrente|  2022-08-21|          14|
| 15|Conta Corrente|  2022-07-15|          15|
| 16|Conta Corrente|  2019-12-27|          16|
| 17|Conta Corrente|  2022-07-31|          17|
| 18|Conta Corrente|  2018-07-13|          18|
| 19|Conta Corrente|  2019-04-14|          19|
| 20|Conta Corrente|  2022-12-07|          20|
+---+--------------+-----------+------------+
only showing top 20 rows
```

In [8]:

```
#Dataframe Cartao
df_cartao_csv = (
    spark.read
    .format('csv')
    .option('header', True)
    .option('delimiter', ';')
    .load(v_diretorio_csv + 'cartao.csv')
)

#Definindo o tipo da coluna
df_cartao_csv_fmt = (
    df_cartao_csv
    .withColumn('id', col('id').cast(IntegerType()))
    .withColumn('id_conta', col('id_conta').cast(IntegerType()))
    .withColumn('id_associado', col('id_associado').cast(IntegerType()))
)

#Criando view do dataframe
df_cartao_csv_fmt.createOrReplaceTempView('cartao_csv')
```

In [9]:

```
df_cartao_csv_fmt.show()
```

```
+---+----------------+----------------+--------+------------+
| id|      num_cartao|     nom_impresso|id_conta|id_associado|
+---+----------------+----------------+--------+------------+
|  1|8692002900010397|   ALÍCIA CARDOSO|       1|           1|
|  2|1360002500020347|    MIRELLA MOURA|       2|           2|
|  3|3935005400035103|RODRIGO FERNANDES|       3|           3|
|  4|4371005900041388|   REBECA CARDOSO|       4|           4|
|  5|9500005100053578|      RAUL BARROS|       5|           5|
|  6|7915003800066514|      JULIA NUNES|       6|           6|
|  7|2184004500079616|      JOÃO MIGUEL|       7|           7|
|  8|2631002700088038|  FRANCISCO GOMES|       8|           8|
|  9|3191005800091087|    VINICIUS LIMA|       9|           9|
```

```
| 10|9897004000108416|    CECÍLIA SOUZA|      10|          10|
| 11|8684005700115334|        ANA JULIA|      11|          11|
| 12|8694004000128933|    ANTHONY NEVES|      12|          12|
| 13|9950003400138288|      LUCAS COSTA|      13|          13|
| 14|4373006600142001|     ANA TEIXEIRA|      14|          14|
| 15|6333007000157004|        JOÃO LUCAS|      15|          15|
| 16|9080006900166160|         BRUNA LUZ|      16|          16|
| 17|6279006700177996|        VITOR HUGO|      17|          17|
| 18|5432003900184311|  SARAH FERNANDES|      18|          18|
| 19|4667007500196222|CECÍLIA RODRIGUES|      19|          19|
| 20|5578004200205193|      NATHAN MOTA|      20|          20|
+---+----------------+-----------------+--------+------------+
only showing top 20 rows
```

In [10]:

```
#Dataframe Movimento
df_movimento_csv = (
    spark.read
    .format("csv")
    .option("header", True)
    .option("delimiter", ";")
    .load(v_diretorio_csv + "movimento.csv")
)

#Definindo o tipo da coluna
df_movimento_csv_fmt = (
    df_movimento_csv
    .withColumn('id', col('id').cast(IntegerType()))
    .withColumn('vlr_transacao', col('vlr_transacao').cast(DecimalType(10,2)))
    .withColumn('data_movimento', col('data_movimento').cast(DateType()))
    .withColumn('id_cartao', col('id_cartao').cast(IntegerType()))
)

#Criando view do dataframe
df_movimento_csv_fmt.createOrReplaceTempView('movimento_csv')
```

In [11]:

```
df_movimento_csv_fmt.show()
```

```
+----+-------------+-----------------+--------------+---------+
|  id|vlr_transacao|    des_transacao|data_movimento|id_cartao|
+----+-------------+-----------------+--------------+---------+
|4249|        65.80|      Restaurante|    2022-05-11|       26|
|4250|        51.64|            Roupa|    2022-05-16|       26|
|4251|       398.16|Posto combustivel|    2022-05-21|       26|
|4252|        55.99|Posto combustivel|    2022-05-25|       26|
|4253|       218.14|         Farmacia|    2022-06-03|       26|
|4254|       543.76|      Restaurante|    2022-06-10|       26|
|4255|       495.44|      Restaurante|    2022-06-13|       26|
|4256|       123.16|         Pet shop|    2022-06-14|       26|
|4257|        35.05|            Roupa|    2022-06-17|       26|
|4258|        23.29|         Pet shop|    2022-06-22|       26|
|4259|       474.36|     Supermercado|    2022-06-23|       26|
|4260|        81.47|      Restaurante|    2022-06-25|       26|
|4261|       464.14|         Pet shop|    2022-07-02|       26|
|4262|       303.69|      Restaurante|    2022-07-04|       26|
|4263|       658.54|      Restaurante|    2022-07-10|       26|
|4264|       376.01|      Restaurante|    2022-07-12|       26|
|4265|       598.44|      Restaurante|    2022-07-13|       26|
|4266|        49.41|     Supermercado|    2022-07-25|       26|
|4267|       493.84|            Roupa|    2022-07-27|       26|
|4268|       499.91|            Roupa|    2022-07-28|       26|
+----+-------------+-----------------+--------------+---------+
only showing top 20 rows
```

In [12]:

```
#Dataframe Encerramento
df_encerramento_csv = (
    spark.read
    .format('csv')
    .option('header', True)
    .option('delimiter', ';')
    .load(v_diretorio_csv + 'encerramento_conta.csv')
)

#Removendo colunas
new_df_encerramento_csv=df_encerramento_csv.drop('semente', 'data_parou_comprar', 'dias_sem_com

#Definindo o tipo da coluna
df_encerramento_csv_fmt = (
    new_df_encerramento_csv
    .withColumn('id', col('id').cast(IntegerType()))
    .withColumn('data_criacao', col('data_criacao').cast(DateType()))
    .withColumn('data_encerramento', col('data_encerramento').cast(DateType()))
)

#Criando view do dataframe
df_encerramento_csv_fmt.createOrReplaceTempView('encerramento_conta_csv')
```

In [13]:

```
df_encerramento_csv_fmt.show()
```

```
+---+------------+-----------------+
| id|data_criacao|data_encerramento|
+---+------------+-----------------+
|  1|  2019-03-28|             null|
|  2|  2021-04-02|             null|
|  3|  2019-05-24|             null|
|  4|  2018-10-22|             null|
|  5|  2022-11-29|             null|
|  6|  2018-05-26|             null|
|  7|  2020-08-23|             null|
|  8|  2019-02-16|             null|
|  9|  2021-03-09|             null|
| 10|  2022-04-09|             null|
| 11|  2019-10-08|             null|
| 12|  2022-04-28|             null|
| 13|  2019-02-15|             null|
| 14|  2022-08-21|             null|
| 15|  2022-07-15|             null|
| 16|  2019-12-27|             null|
| 17|  2022-07-31|             null|
| 18|  2018-07-13|             null|
| 19|  2019-04-14|             null|
| 20|  2022-12-07|             null|
+---+------------+-----------------+
only showing top 20 rows
```

In [14]:

```
#Dataframe Fatura
df_fatura_csv = (
    spark.read
    .format('csv')
    .option('header', True)
    .option('delimiter', ';')
    .load(v_diretorio_csv + 'fatura.csv')
)

#Definindo o tipo da coluna
df_fatura_csv_fmt = (
    df_fatura_csv
    .withColumn('id', col('id').cast(IntegerType()))
```

```
    .withColumn('data_vencimento_fatura', col('data_vencimento_fatura').cast(DateType()))
    .withColumn('vlr_fatura', col('vlr_fatura').cast(DecimalType(10,2)))
    .withColumn('data_pagamento_fatura', col('data_pagamento_fatura').cast(DateType()))
    .withColumn('qtd_dias_atraso_pgto', col('qtd_dias_atraso_pgto').cast(IntegerType()))
    .withColumn('id_cartao', col('id_cartao').cast(IntegerType()))
)

#Criando view do dataframe
df_fatura_csv_fmt.createOrReplaceTempView('fatura_csv')
```

In [15]:

```
df_fatura_csv_fmt.show()
```

```
+---+---------------------+---------+--------------------+-------------------+---------+
| id|data_vencimento_fatura|vlr_fatura|data_pagamento_fatura|qtd_dias_atraso_pgto|id_cartao|
+---+---------------------+---------+--------------------+-------------------+---------+
|  1|           2019-03-15|      0.00|          2019-03-15|                  0|        1|
|  2|           2019-04-15|   1470.86|          2019-04-11|                  0|        1|
|  3|           2019-05-15|   1634.88|          2019-05-11|                  0|        1|
|  4|           2019-06-15|    437.91|          2019-06-11|                  0|        1|
|  5|           2019-07-15|   1006.45|          2019-07-14|                  0|        1|
|  6|           2019-08-15|    932.13|          2019-08-12|                  0|        1|
|  7|           2019-09-15|    693.69|          2019-09-12|                  0|        1|
|  8|           2019-10-15|    349.34|          2019-10-12|                  0|        1|
|  9|           2019-11-15|   1609.69|          2019-11-11|                  0|        1|
| 10|           2019-12-15|   1456.59|          2019-12-14|                  0|        1|
| 11|           2020-01-15|   1129.89|          2020-01-14|                  0|        1|
| 12|           2020-02-15|    186.44|          2020-02-15|                  0|        1|
| 13|           2020-03-15|   1350.80|          2020-03-11|                  0|        1|
| 14|           2020-04-15|    880.98|          2020-04-12|                  0|        1|
| 15|           2020-05-15|   1493.61|          2020-05-12|                  0|        1|
| 16|           2020-06-15|    518.68|          2020-06-10|                  0|        1|
| 17|           2020-07-15|    958.56|          2020-07-11|                  0|        1|
| 18|           2020-08-15|    210.13|          2020-08-15|                  0|        1|
| 19|           2020-09-15|   1059.57|          2020-09-11|                  0|        1|
| 20|           2020-10-15|   1224.60|          2020-10-13|                  0|        1|
+---+---------------------+---------+--------------------+-------------------+---------+
only showing top 20 rows
```

Carregando dataframes e views com os dados do banco de dados do target

In [16]:

```
#Carregando dados no Dataframe
df_associado_tgt = (
    spark.read
    .format('jdbc')
    .option('url', v_url_jdbc)
    .option('dbtable', 'target.associado')
    .option('user', v_user_jdbc)
    .option('password', v_pass_jdbc)
    .load()
)

#Criando view do dataframe
df_associado_tgt.createOrReplaceTempView('associado_tgt')
```

In [17]:

```
#Carregando dados no Dataframe
df_conta_tgt = (
    spark.read
    .format('jdbc')
    .option('url', v_url_jdbc)
    .option('dbtable', 'target.conta')
    .option('user', v_user_jdbc)
    .option('password', v_pass_jdbc)
    .load()
)
```

```
#Criando view do dataframe
df_conta_tgt.createOrReplaceTempView('conta_tgt')
```

In [18]:

```
#Carregando dados no Dataframe
df_cartao_tgt = (
    spark.read
    .format('jdbc')
    .option('url', v_url_jdbc)
    .option('dbtable', 'target.cartao')
    .option('user', v_user_jdbc)
    .option('password', v_pass_jdbc)
    .load()
)

#Criando view do dataframe
df_cartao_tgt.createOrReplaceTempView('cartao_tgt')
```

In [19]:

```
#Carregando dados no Dataframe
df_movimento_tgt = (
    spark.read
    .format('jdbc')
    .option('url', v_url_jdbc)
    .option('dbtable', 'target.movimento')
    .option('user', v_user_jdbc)
    .option('password', v_pass_jdbc)
    .load()
)

#Criando view do dataframe
df_movimento_tgt.createOrReplaceTempView('movimento_tgt')
```

In [20]:

```
#Carregando dados no Dataframe
df_encerramento_conta_tgt = (
    spark.read
    .format('jdbc')
    .option('url', v_url_jdbc)
    .option('dbtable', 'target.encerramento_conta')
    .option('user', v_user_jdbc)
    .option('password', v_pass_jdbc)
    .load()
)

#Criando view do dataframe
df_encerramento_conta_tgt.createOrReplaceTempView('encerramento_conta_tgt')
```

In [21]:

```
#Carregando dados no Dataframe
df_fatura_tgt = (
    spark.read
    .format('jdbc')
    .option('url', v_url_jdbc)
    .option('dbtable', 'target.fatura')
    .option('user', v_user_jdbc)
    .option('password', v_pass_jdbc)
    .load()
)

#Criando view do dataframe
df_fatura_tgt.createOrReplaceTempView('fatura_tgt')
```

Teste unitário

O objetivo do teste unitário, é validar se todos os dados do arquivo csv foram carregados corretamente.

Arquivo associado

In [22]:

```
df_associado_validacao=spark.sql('''
    select
        csv.id as id_csv,
        tgt.id as id_tgt,
        csv.nome as nome_csv,
        tgt.nome as nome_tgt,
        csv.sobrenome as sobrenome_csv,
        tgt.sobrenome as sobrenome_tgt,
        csv.idade as idade_csv,
        tgt.idade as idade_tgt,
        csv.email as email_csv,
        tgt.email as email_tgt

    from associado_csv csv

    inner join associado_tgt tgt
    on tgt.id=csv.id
''')

qtd_associados_csv=df_associado_csv.count()
qtd_associados_encontrados=df_associado_validacao.count()

print("Validação da quantidade de registros: CSV x Target (Data Lake)")
print()
print(f"Qtd. registros do arquivo csv: {qtd_associados_csv}.")
print(f"Qtd. registros encontrados no target: {qtd_associados_encontrados}. ")
print()
print("Comparação de 20 registros aleatórios")
df_associado_validacao.sample(False, 0.1, seed=0).limit(20).show()
```

```
Validação da quantidade de registros: CSV x Target (Data Lake)

Qtd. registros do arquivo csv: 9951.
Qtd. registros encontrados no target: 9951.

Comparação de 20 registros aleatórios
+------+------+---------+---------+-------------+-------------+---------+---------+-----------
--------+--------------------+
|id_csv|id_tgt| nome_csv| nome_tgt|sobrenome_csv|sobrenome_tgt|idade_csv|idade_tgt|          e
mail_csv|           email_tgt|
+------+------+---------+---------+-------------+-------------+---------+---------+-----------
--------+--------------------+
|     3|     3|  Rodrigo|  Rodrigo|    Fernandes|    Fernandes|       54|       54|rodrigo.fern
andes...|rodrigo.fernandes...|
|    10|    10|  Cecília|  Cecília|        Souza|        Souza|       40|       40|cecília.souz
a@uol...|cecília.souza@uol...|
|    17|    17|    Vitor|    Vitor|         Hugo|         Hugo|       67|       67|vitor.hugo@h
otmai...|vitor.hugo@hotmai...|
|    27|    27|      Ana|      Ana|        Luiza|        Luiza|       27|       27| ana.luiza@g
mail.com| ana.luiza@gmail.com|
|    47|    47|    Diogo|    Diogo|        Cunha|        Cunha|       70|       70|diogo.cunha@
terra...|diogo.cunha@terra...|
|    58|    58|   Thiago|   Thiago|    Rodrigues|    Rodrigues|       37|       37|thiago.rodri
gues@...|thiago.rodrigues@...|
|    59|    59|     João|     João|        Lucas|        Lucas|       40|       40|joão.lucas@g
mail.com|joão.lucas@gmail.com|
|    67|    67|Gabrielly|Gabrielly|      Ribeiro|      Ribeiro|       48|       48|gabrielly.ri
beiro...|gabrielly.ribeiro...|
|    80|    80|   Esther|   Esther|          Luz|          Luz|       66|       66|esther.luz@y
ahoo....|esther.luz@yahoo....|
|    93|    93|  Cecília|  Cecília|     Silveira|     Silveira|       23|       23|cecília.silv
eira@...|cecília.silveira@...|
|   115|   115|Gabrielly|Gabrielly|        Neves|        Neves|       51|       51|gabrielly.ne
ves@t...|gabrielly.neves@t...|
|   127|   127|    Diego|    Diego|        Souza|        Souza|       45|       45|diego.souza@
```

```
gmail...|diego.souza@gmail...|
|   139|   139|      Ana|      Ana|        Clara|        Clara|       25|       25|ana.clara@te
rra.c...|ana.clara@terra.c...|
|   141|   141|    Srta.|    Srta.|          Ana|          Ana|       31|       31|srta..ana@ho
tmail...|srta..ana@hotmail...|
|   146|   146|  Anthony|  Anthony|        Sales|        Sales|       63|       63|anthony.sale
s@hot...|anthony.sales@hot...|
|   149|   149|   Sophia|   Sophia|       Campos|       Campos|       74|       74|sophia.campo
s@yah...|sophia.campos@yah...|
|   171|   171|     João|     João|    Guilherme|    Guilherme|       20|       20|joão.guilher
me@ya...|joão.guilherme@ya...|
|   178|   178|   Daniel|   Daniel|     Silveira|     Silveira|       29|       29|daniel.silve
ira@g...|daniel.silveira@g...|
|   182|   182|     Raul|     Raul|      Moreira|      Moreira|       23|       23|raul.moreira
@gmai...|raul.moreira@gmai...|
|   185|   185|   Emilly|   Emilly|       Barros|       Barros|       72|       72|emilly.barro
s@hot...|emilly.barros@hot...|
+------+------+---------+---------+-------------+-------------+---------+---------+-----------
--------+--------------------+
```

Arquivo conta

In [23]:

```
df_conta_validacao=spark.sql('''
    select
         csv.id as id_csv,
         tgt.id as id_tgt,
         csv.tipo as tipo_csv,
         tgt.tipo as tipo_tgt,
         csv.data_criacao as data_criacao_csv,
         tgt.data_criacao as data_criacao_tgt,
         csv.id_associado as id_associado_csv,
         tgt.id_associado as id_associado_tgt

    from conta_csv csv

    inner join conta_tgt tgt
    on tgt.id=csv.id
''')

qtd_conta_csv=df_conta_csv.count()
qtd_conta_encontrados=df_conta_validacao.count()

print("Validação da quantidade de registros: CSV x Target (Data Lake)")
print()
print(f"Qtd. registros do arquivo csv: {qtd_conta_csv}.")
print(f"Qtd. registros encontrados no target: {qtd_conta_encontrados}. ")
print()
print("Comparação de 20 registros aleatórios")
df_conta_validacao.sample(False, 0.1, seed=0).limit(20).show()
```

```
Validação da quantidade de registros: CSV x Target (Data Lake)

Qtd. registros do arquivo csv: 10001.
Qtd. registros encontrados no target: 10001.

Comparação de 20 registros aleatórios
+------+------+--------------+--------------+---------------+---------------+---------------
+----------------+
|id_csv|id_tgt|      tipo_csv|      tipo_tgt|data_criacao_csv|data_criacao_tgt|id_associado_csv
|id_associado_tgt|
+------+------+--------------+--------------+---------------+---------------+---------------
+----------------+
|   471|   471|Conta Corrente|Conta Corrente|     2018-04-12|     2018-04-12|            471
|             471|
|  1591|  1591|Conta Corrente|Conta Corrente|     2021-02-01|     2021-02-01|           1591
|            1591|
|  2659|  2659|Conta Corrente|Conta Corrente|     2018-05-26|     2018-05-26|           2659
```

```
|           2659|
|  4900|  4900|Conta Corrente|Conta Corrente|      2021-10-24|      2021-10-24|           4900
|           4900|
|  7982|  7982|Conta Corrente|Conta Corrente|      2021-07-14|      2021-07-14|           7982
|           7982|
|   243|   243|Conta Corrente|Conta Corrente|      2022-02-09|      2022-02-09|            243
|            243|
|   392|   392|Conta Corrente|Conta Corrente|      2020-07-30|      2020-07-30|            392
|             -1|
|  1127|  1127|Conta Corrente|Conta Corrente|      2020-03-16|      2020-03-16|           1127
|           1127|
|  2811|  2811|Conta Corrente|Conta Corrente|      2020-04-01|      2020-04-01|           2811
|           2811|
|  6623|  6623|Conta Corrente|Conta Corrente|      2020-08-05|      2020-08-05|           6623
|           6623|
|  1699|  1699|Conta Corrente|Conta Corrente|      2018-01-21|      2018-01-21|           1699
|           1699|
|  3704|  3704|Conta Corrente|Conta Corrente|      2021-06-09|      2021-06-09|           3704
|           3704|
|  6559|  6559|Conta Corrente|Conta Corrente|      2022-07-27|      2022-07-27|           6559
|           6559|
|  6825|  6825|Conta Corrente|Conta Corrente|      2021-04-02|      2021-04-02|           6825
|           6825|
|  8222|  8222|Conta Corrente|Conta Corrente|      2022-10-05|      2022-10-05|           8222
|           8222|
|  8924|  8924|Conta Corrente|Conta Corrente|      2022-03-12|      2022-03-12|           8924
|           8924|
|  2748|  2748|Conta Corrente|Conta Corrente|      2019-05-03|      2019-05-03|           2748
|           2748|
|  3876|  3876|Conta Corrente|Conta Corrente|      2022-08-09|      2022-08-09|           3876
|           3876|
|  4186|  4186|Conta Corrente|Conta Corrente|      2022-04-20|      2022-04-20|           4186
|           4186|
|  4684|  4684|Conta Corrente|Conta Corrente|      2019-10-07|      2019-10-07|           4684
|           4684|
+------+------+--------------+--------------+----------------+----------------+----------------
+----------------+
```

Arquivo cartão

In [24]:

```
df_cartao_validacao=spark.sql('''
    select
          csv.id as id_csv,
          tgt.id as id_tgt,
          csv.num_cartao as num_cartao_csv,
          tgt.num_cartao as num_cartao_tgt,
          csv.nom_impresso as nom_impresso_csv,
          tgt.nom_impresso as nom_impresso_tgt,
          csv.id_conta as id_conta_csv,
          tgt.id_conta as id_conta_tgt,
          csv.id_associado as id_associado_csv,
          tgt.id_associado as id_associado_tgt

    from cartao_csv csv

    inner join cartao_tgt tgt
    on tgt.id=csv.id
''')

qtd_cartao_csv=df_cartao_csv.count()
qtd_cartao_encontrados=df_cartao_validacao.count()

print("Validação da quantidade de registros: CSV x Target (Data Lake)")
print()
print(f"Qtd. registros do arquivo csv: {qtd_cartao_csv}.")
print(f"Qtd. registros encontrados no target: {qtd_cartao_encontrados}. ")
print()
```

```
print("Comparação de 20 registros aleatórios")
df_cartao_validacao.sample(False, 0.1, seed=0).limit(20).show()
```

```
Validação da quantidade de registros: CSV x Target (Data Lake)

Qtd. registros do arquivo csv: 10001.
Qtd. registros encontrados no target: 10001.

Comparação de 20 registros aleatórios
+------+------+---------------+---------------+---------------+---------------+-----------
+-----------+---------------+---------------+
|id_csv|id_tgt|  num_cartao_csv|  num_cartao_tgt|nom_impresso_csv|nom_impresso_tgt|id_conta_csv
|id_conta_tgt|id_associado_csv|id_associado_tgt|
+------+------+---------------+---------------+---------------+---------------+-----------
+-----------+---------------+---------------+
|   471|   471|5803004704713590|5803004704713590| MIRELLA BARBOSA| MIRELLA BARBOSA|         471
|         471|             471|             471|
|  1591|  1591|7542003915911574|7542003915911574|     LUCCA COSTA|     LUCCA COSTA|        1591
|        1591|            1591|            1591|
|  2659|  2659|4877006326597672|4877006326597672|   DANIELA CUNHA|   DANIELA CUNHA|        2659
|        2659|            2659|            2659|
|  4900|  4900|2674005249002878|2674005249002878|    YURI ALMEIDA|    YURI ALMEIDA|        4900
|        4900|            4900|            4900|
|  7982|  7982|5888003179825010|5888003179825010|   MARIANA PINTO|   MARIANA PINTO|        7982
|        7982|            7982|            7982|
|   243|   243|4745003402435264|4745003402435264|    LUCAS NOVAES|    LUCAS NOVAES|         243
|         243|             243|             243|
|   392|   392|7516002203922622|7516002203922622|      PAULO ROSA|      PAULO ROSA|         392
|         392|             392|              -1|
|  1127|  1127|9893005011276433|9893005011276433|    ANA CAROLINA|    ANA CAROLINA|        1127
|        1127|            1127|            1127|
|  2811|  2811|8598006628116253|8598006628116253|    HELENA PORTO|    HELENA PORTO|        2811
|        2811|            2811|            2811|
|  6623|  6623|2702004266238648|2702004266238648|     JOÃO MIGUEL|     JOÃO MIGUEL|        6623
|        6623|            6623|            6623|
|  1650|  1650|2352005716504428|2352005716504428|       ANA JÚLIA|       ANA JÚLIA|        1650
|        1650|            1650|            1650|
|  3488|  3488|5796007034886987|5796007034886987|    MIGUEL SOUZA|    MIGUEL SOUZA|        3488
|        3488|            3488|            3488|
|  6482|  6482|2440003464828713|2440003464828713|    ENZO GABRIEL|    ENZO GABRIEL|        6482
|        6482|            6482|            6482|
|  6622|  6622|6723005566221778|6723005566221778| MARIANA CARDOSO| MARIANA CARDOSO|        6622
|        6622|            6622|            6622|
|  7879|  7879|4764001878794303|4764001878794303| NICOLAS BARBOSA| NICOLAS BARBOSA|        7879
|        7879|            7879|            7879|
|  8407|  8407|4300004584077301|4300004584077301|       DAVI LUIZ|       DAVI LUIZ|        8407
|        8407|            8407|            8407|
|  2721|  2721|2277003927218254|2277003927218254|   HELOÍSA VIANA|   HELOÍSA VIANA|        2721
|        2721|            2721|            2721|
|  3796|  3796|4517003237964041|4517003237964041|    DANIEL PIRES|    DANIEL PIRES|        3796
|        3796|            3796|            3796|
|  4078|  4078|2421002340787656|2421002340787656|      SRTA. LAÍS|      SRTA. LAÍS|        4078
|        4078|            4078|            4078|
|  4364|  4364|1486007043646939|1486007043646939|      NINA ROCHA|      NINA ROCHA|        4364
|        4364|            4364|            4364|
+------+------+---------------+---------------+---------------+---------------+-----------
+-----------+---------------+---------------+
```

Arquivo movimento

In [25]:

```
df_movimento_validacao=spark.sql('''
    select
        csv.id as id_csv,
        tgt.id as id_tgt,
        csv.vlr_transacao as vlr_transacao_csv,
        tgt.vlr_transacao as vlr_transacao_tgt,
        csv.des_transacao as des_transacao_csv,
        tgt.des_transacao as des_transacao_tgt,
```

```
        csv.data_movimento as data_movimento_csv,
        tgt.data_movimento as data_movimento_tgt,
        csv.id_cartao as id_cartao_csv,
        tgt.id_cartao as id_cartao_tgt

    from movimento_csv csv

    inner join movimento_tgt tgt
    on tgt.id=csv.id
''')

qtd_movimento_csv=df_movimento_csv.count()
qtd_movimento_encontrados=df_movimento_validacao.count()

print("Validação da quantidade de registros: CSV x Target (Data Lake)")
print()
print(f"Qtd. registros do arquivo csv: {qtd_movimento_csv}.")
print(f"Qtd. registros encontrados no target: {qtd_movimento_encontrados}. ")
print()
print("Comparação de 20 registros aleatórios")
df_movimento_validacao.sample(False, 0.1, seed=0).limit(20).show()
```

```
Validação da quantidade de registros: CSV x Target (Data Lake)

Qtd. registros do arquivo csv: 1617874.
Qtd. registros encontrados no target: 1617874.

Comparação de 20 registros aleatórios
[Stage 43:=============================================>       (174 + 1) / 200]
+------+------+----------------+----------------+----------------+----------------+--------
----------+------------------+-------------+-------------+
|id_csv|id_tgt|vlr_transacao_csv|vlr_transacao_tgt|des_transacao_csv|des_transacao_tgt|data_mov
imento_csv|data_movimento_tgt|id_cartao_csv|id_cartao_tgt|
+------+------+----------------+----------------+----------------+----------------+--------
----------+------------------+-------------+-------------+
|   471|   471|           288.77|           288.77|Posto combustivel|Posto combustivel|
2020-12-02|        2020-12-02|            3|            3|
|  1591|  1591|           126.65|           126.65|      Restaurante|      Restaurante|
2021-04-19|        2021-04-19|            8|            8|
|  2659|  2659|            23.54|            23.54|     Supermercado|     Supermercado|
2021-10-06|        2021-10-06|           16|           16|
|  4900|  4900|            29.42|            29.42|     Supermercado|     Supermercado|
2018-12-11|        2018-12-11|           31|           31|
|  7982|  7982|           377.97|           377.97|      Restaurante|      Restaurante|
2022-01-12|        2022-01-12|           47|           47|
| 10206| 10206|           410.86|           410.86|         Pet shop|         Pet shop|
2019-10-11|        2019-10-11|           67|           67|
| 10362| 10362|           233.82|           233.82|         Farmacia|         Farmacia|
2022-02-07|        2022-02-07|           67|           67|
| 11858| 11858|            95.82|            95.82|      Restaurante|      Restaurante|
2019-11-09|        2019-11-09|           78|           78|
| 15447| 15447|           366.24|           366.24|Posto combustivel|Posto combustivel|
2022-01-21|        2022-01-21|          101|          101|
| 17679| 17679|           318.61|           318.61|     Supermercado|     Supermercado|
2021-08-14|        2021-08-14|          112|          112|
| 21220| 21220|            91.25|            91.25|            Roupa|            Roupa|
2020-02-15|        2020-02-15|          138|          138|
| 24171| 24171|           471.44|           471.44|     Supermercado|     Supermercado|
2021-05-10|        2021-05-10|          151|          151|
| 26708| 26708|           297.04|           297.04|      Restaurante|      Restaurante|
2021-02-06|        2021-02-06|          164|          164|
| 27484| 27484|           299.51|           299.51|         Pet shop|         Pet shop|
2020-03-04|        2020-03-04|          168|          168|
| 28124| 28124|           133.88|           133.88|      Restaurante|      Restaurante|
2019-11-19|        2019-11-19|          172|          172|
| 28577| 28577|            93.63|            93.63|         Farmacia|         Farmacia|
2019-09-10|        2019-09-10|          175|          175|
```

```
| 31367| 31367|           273.98|           273.98|     Supermercado|     Supermercado|
2019-12-06|        2019-12-06|          188|          188|
| 32445| 32445|           118.39|           118.39|         Pet shop|         Pet shop|
2022-05-22|        2022-05-22|          193|          193|
| 32855| 32855|           183.57|           183.57|      Restaurante|      Restaurante|
2021-02-03|        2021-02-03|          195|          195|
| 33569| 33569|           406.60|           406.60|         Pet shop|         Pet shop|
2022-09-26|        2022-09-26|          199|          199|
+------+------+-----------------+-----------------+-----------------+-----------------+--------
----------+------------------+-------------+-------------+
```

Arquivo encerramento conta

In [26]:

```
df_encerramento_conta_validacao=spark.sql('''
    select
        csv.id as id_csv,
        tgt.id as id_tgt,
        csv.data_criacao as data_criacao_csv,
        tgt.data_criacao as data_criacao_tgt,
        csv.data_encerramento as data_encerramento_csv,
        tgt.data_encerramento as data_encerramento_tgt

    from encerramento_conta_csv csv

    inner join encerramento_conta_tgt tgt
    on tgt.id=csv.id
''')

qtd_encerramento_conta_csv=df_encerramento_csv.count()
qtd_encerramento_conta_encontrados=df_encerramento_conta_validacao.count()

print("Validação da quantidade de registros: CSV x Target (Data Lake)")
print()
print(f"Qtd. registros do arquivo csv: {qtd_encerramento_conta_csv}.")
print(f"Qtd. registros encontrados no target: {qtd_encerramento_conta_encontrados}. ")
print()
print("Comparação de 20 registros aleatórios")
df_encerramento_conta_validacao.sample(False, 0.1, seed=0).limit(20).show()
```

```
Validação da quantidade de registros: CSV x Target (Data Lake)

Qtd. registros do arquivo csv: 10000.
Qtd. registros encontrados no target: 10000.

Comparação de 20 registros aleatórios
+------+------+---------------+---------------+--------------------+--------------------+
|id_csv|id_tgt|data_criacao_csv|data_criacao_tgt|data_encerramento_csv|data_encerramento_tgt|
+------+------+---------------+---------------+--------------------+--------------------+
|   471|   471|     2018-04-12|     2018-04-12|                null|                null|
|  1591|  1591|     2021-02-01|     2021-02-01|                null|                null|
|  2659|  2659|     2018-05-26|     2018-05-26|                null|                null|
|  4900|  4900|     2021-10-24|     2021-10-24|                null|                null|
|  7982|  7982|     2021-07-14|     2021-07-14|          2022-10-18|          2022-10-18|
|   243|   243|     2022-02-09|     2022-02-09|                null|                null|
|   392|   392|     2020-07-30|     2020-07-30|                null|                null|
|  1127|  1127|     2020-03-16|     2020-03-16|                null|                null|
|  2811|  2811|     2020-04-01|     2020-04-01|                null|                null|
|  6623|  6623|     2020-08-05|     2020-08-05|                null|                null|
|  1699|  1699|     2018-01-21|     2018-01-21|                null|                null|
|  3704|  3704|     2021-06-09|     2021-06-09|                null|                null|
|  6559|  6559|     2022-07-27|     2022-07-27|                null|                null|
|  6825|  6825|     2021-04-02|     2021-04-02|                null|                null|
|  8222|  8222|     2022-10-05|     2022-10-05|                null|                null|
|  8924|  8924|     2022-03-12|     2022-03-12|                null|                null|
|  2748|  2748|     2019-05-03|     2019-05-03|                null|                null|
|  3876|  3876|     2022-08-09|     2022-08-09|                null|                null|
```

```
|  4186|  4186|      2022-04-20|      2022-04-20|                null|                null|
|  4684|  4684|      2019-10-07|      2019-10-07|                null|                null|
+------+------+----------------+----------------+--------------------+--------------------+
```

Arquivo fatura

In [27]:

```
df_fatura_validacao=spark.sql('''
    select
        csv.id as id_csv,
        tgt.id as id_tgt,
        csv.data_vencimento_fatura as data_vencimento_fatura_csv,
        tgt.data_vencimento_fatura as data_vencimento_fatura_tgt,
        csv.vlr_fatura as vlr_fatura_csv,
        tgt.vlr_fatura as vlr_fatura_tgt,
        csv.data_pagamento_fatura as data_pagamento_fatura_csv,
        tgt.data_pagamento_fatura as data_pagamento_fatura_tgt,
        csv.qtd_dias_atraso_pgto as qtd_dias_atraso_pgto_csv,
        tgt.qtd_dias_atraso_pgto as qtd_dias_atraso_pgto_tgt,
        csv.id_cartao as id_cartao_csv,
        tgt.id_cartao as id_cartao_tgt

    from fatura_csv csv

    inner join fatura_tgt tgt
    on tgt.id=csv.id
''')

qtd_fatura_csv=df_fatura_csv.count()
qtd_fatura_encontrados=df_fatura_validacao.count()

print("Validação da quantidade de registros: CSV x Target (Data Lake)")
print()
print(f"Qtd. registros do arquivo csv: {qtd_fatura_csv}.")
print(f"Qtd. registros encontrados no target: {qtd_fatura_encontrados}. ")
print()
print("Comparação de 20 registros aleatórios")
df_fatura_validacao.sample(False, 0.1, seed=0).limit(20).show()
```

```
Validação da quantidade de registros: CSV x Target (Data Lake)

Qtd. registros do arquivo csv: 301613.
Qtd. registros encontrados no target: 301613.

Comparação de 20 registros aleatórios

+------+------+-------------------------+-------------------------+-------------+-----------
---+-------------------------+-------------------------+-------------------------+-------------
----------+-------------+-------------+
|id_csv|id_tgt|data_vencimento_fatura_csv|data_vencimento_fatura_tgt|vlr_fatura_csv|vlr_fatura_
tgt|data_pagamento_fatura_csv|data_pagamento_fatura_tgt|qtd_dias_atraso_pgto_csv|qtd_dias_atras
o_pgto_tgt|id_cartao_csv|id_cartao_tgt|
+------+------+-------------------------+-------------------------+-------------+-----------
---+-------------------------+-------------------------+-------------------------+-------------
----------+-------------+-------------+
|   471|   471|               2022-07-15|               2022-07-15|         0.00|
0.00|              2022-07-15|              2022-07-15|                        0|
0|           17|           17|
|  1591|  1591|               2020-11-15|               2020-11-15|        57.62|         5
7.62|              2020-11-15|              2020-11-15|                        0|
0|           56|           56|
|  2659|  2659|               2021-04-15|               2021-04-15|      1226.84|       122
6.84|              2021-04-11|              2021-04-11|                        0|
0|           96|           96|
|  4900|  4900|               2021-09-15|               2021-09-15|      2132.80|       213
2.80|              2021-09-10|              2021-09-10|                        0|
0|          161|          161|
|  7982|  7982|               2020-01-15|               2020-01-15|       549.18|        54
```

```
9.18|              2020-01-15|              2020-01-15|                   0|
0|          261|          261|
| 10206| 10206|              2020-02-15|              2020-02-15|         83.15|          8
3.15|              2020-02-15|              2020-02-15|                   0|
0|          336|          336|
| 10362| 10362|              2019-06-15|              2019-06-15|         39.43|          3
9.43|              2019-06-15|              2019-06-15|                   0|
0|          341|          341|
| 11858| 11858|              2021-06-15|              2021-06-15|        245.73|         24
5.73|              2021-06-15|              2021-06-15|                   0|
0|          390|          390|
| 15447| 15447|              2022-10-15|              2022-10-15|          0.00|
0.00|              2022-10-15|              2022-10-15|                   0|
0|          501|          501|
| 17679| 17679|              2020-10-15|              2020-10-15|       3102.76|        310
2.76|              2020-10-10|              2020-10-10|                   0|
0|          582|          582|
| 21220| 21220|              2018-12-15|              2018-12-15|       2322.56|        232
2.56|              2018-12-10|              2018-12-10|                   0|
0|          691|          691|
| 24171| 24171|              2020-09-15|              2020-09-15|       1751.33|        175
1.33|              2020-09-10|              2020-09-10|                   0|
0|          783|          783|
| 26708| 26708|              2020-01-15|              2020-01-15|        664.39|         66
4.39|              2020-01-15|              2020-01-15|                   0|
0|          870|          870|
| 27484| 27484|              2022-12-15|              2022-12-15|        642.44|         64
2.44|              2022-12-15|              2022-12-15|                   0|
0|          891|          891|
| 28124| 28124|              2022-06-15|              2022-06-15|       2604.59|        260
4.59|              2022-06-10|              2022-06-10|                   0|
0|          912|          912|
| 28577| 28577|              2021-01-15|              2021-01-15|       3121.20|        312
1.20|              2021-01-11|              2021-01-11|                   0|
0|          925|          925|
| 31367| 31367|              2022-02-15|              2022-02-15|       1404.13|        140
4.13|              2022-02-13|              2022-02-13|                   0|
0|         1015|         1015|
| 32445| 32445|              2022-12-15|              2022-12-15|        869.61|         86
9.61|              2022-12-12|              2022-12-12|                   0|
0|         1052|         1052|
| 32855| 32855|              2022-04-15|              2022-04-15|          0.00|
0.00|              2022-04-15|              2022-04-15|                   0|
0|         1064|         1064|
| 33569| 33569|              2021-12-15|              2021-12-15|       1516.94|        151
6.94|              2021-12-10|              2021-12-10|                   0|
0|         1090|         1090|
+------+------+-------------------------+-------------------------+-------------+----------
---+-------------------------+-------------------------+-----------------------+-------------
----------+-------------+-------------+
```

Teste integrado

O objetivo do teste integrado é validar se existe algum problema na integridade de relacionamento entre os dados.

Validar a integridade de relacionamento entre as tabelas Conta e Associado

In [28]:

```
df_nao_encontrado=spark.sql('''
    select tgt1.id_associado

    from conta_tgt tgt1

    left join associado_tgt tgt2
    on tgt2.id=tgt1.id_associado

    where tgt2.id is null
```

```
''')

qtd_nao_encontrado=df_nao_encontrado.count()

df_tratado=spark.sql('''
    select tgt1.*

    from conta_tgt tgt1

    where tgt1.id_associado=-1
    and tgt1.id<>-1
''')

qtd_tratados=df_tratado.count()

print("Validação da quantidade de registros: CSV x Target (Data Lake)")
print()
print(f"Qtd. registros não encontrados por problemas de integridade: {qtd_nao_encontrado}.")
print(f"Qtd. registros tratados com '-1': {qtd_tratados}")
print()
print("20 registros aleatórios com tratamento de: -1")
df_tratado.sample(False, 0.99, seed=0).limit(20).show()
```

```
Validação da quantidade de registros: CSV x Target (Data Lake)

Qtd. registros não encontrados por problemas de integridade: 0.
Qtd. registros tratados com '-1': 50

20 registros aleatórios com tratamento de: -1
+----+--------------+------------+------------+
|  id|          tipo|data_criacao|id_associado|
+----+--------------+------------+------------+
| 392|Conta Corrente|  2020-07-30|          -1|
|6011|Conta Corrente|  2020-07-20|          -1|
|1290|Conta Corrente|  2018-10-10|          -1|
|4823|Conta Corrente|  2019-08-19|          -1|
|4481|Conta Corrente|  2022-12-09|          -1|
|7519|Conta Corrente|  2019-12-24|          -1|
|1266|Conta Corrente|  2021-01-28|          -1|
|8818|Conta Corrente|  2019-03-07|          -1|
|9176|Conta Corrente|  2022-10-07|          -1|
|4352|Conta Corrente|  2018-01-24|          -1|
|7581|Conta Corrente|  2019-12-20|          -1|
|7158|Conta Corrente|  2018-02-22|          -1|
|6748|Conta Corrente|  2021-12-28|          -1|
|9350|Conta Corrente|  2021-04-02|          -1|
|1967|Conta Corrente|  2018-06-19|          -1|
|6312|Conta Corrente|  2021-07-03|          -1|
|4071|Conta Corrente|  2018-08-27|          -1|
|2674|Conta Corrente|  2018-12-16|          -1|
|2258|Conta Corrente|  2020-06-23|          -1|
|2827|Conta Corrente|  2021-09-24|          -1|
+----+--------------+------------+------------+
```

Validar a integridade de relacionamento entre as tabelas Cartao, Conta e Associado

In [29]:

```
df_nao_encontrado_1=spark.sql('''
    select tgt1.id_associado

    from cartao_tgt tgt1

    left join associado_tgt tgt2
    on tgt2.id=tgt1.id_associado

    where tgt2.id is null
''')
```

```
qtd_nao_encontrado_1=df_nao_encontrado_1.count()

df_nao_encontrado_2=spark.sql('''
    select tgt1.id_associado

    from cartao_tgt tgt1

    left join conta_tgt tgt2
    on tgt2.id=tgt1.id_conta

    where tgt2.id is null
''')
qtd_nao_encontrado_2=df_nao_encontrado_2.count()

df_tratado_1=spark.sql('''
    select tgt1.*

    from cartao_tgt tgt1

    where tgt1.id_associado=-1
    and tgt1.id<>-1
''')

qtd_tratados_1=df_tratado_1.count()

df_tratado_2=spark.sql('''
    select tgt1.*

    from cartao_tgt tgt1

    where tgt1.id_conta=-1
    and tgt1.id<>-1
''')

qtd_tratados_2=df_tratado_2.count()

print("Validação da quantidade de registros: CSV x Target (Data Lake)")
print()
print(f"Qtd. registros não encontrados por problemas de integridade (Associado): {qtd_nao_encor
print(f"Qtd. registros não encontrados por problemas de integridade (Conta): {qtd_nao_encontrac
print()
print(f"Qtd. registros tratados com '-1' (Associado): {qtd_tratados_1}")
print("20 registros aleatórios com tratamento de: -1 (Associado)")
df_tratado_1.sample(False, 0.99, seed=0).limit(20).show()
print()
print(f"Qtd. registros tratados com '-1'(Conta): {qtd_tratados_2}")
print("20 registros aleatórios com tratamento de: -1 (Conta)")
df_tratado_2.show()
```

```
Validação da quantidade de registros: CSV x Target (Data Lake)

Qtd. registros não encontrados por problemas de integridade (Associado): 0.
Qtd. registros não encontrados por problemas de integridade (Conta): 0.

Qtd. registros tratados com '-1' (Associado): 50
20 registros aleatórios com tratamento de: -1 (Associado)
+----+----------------+---------------+--------+------------+
|  id|      num_cartao|   nom_impresso|id_conta|id_associado|
+----+----------------+---------------+--------+------------+
| 392|7516002203922622|     PAULO ROSA|     392|          -1|
|6011|7080006960112516|   JOÃO GABRIEL|    6011|          -1|
|1290|3562005212902320|  MIRELLA ROCHA|    1290|          -1|
|4823|1721005648239215|    JOÃO MIGUEL|    4823|          -1|
|4481|2690003944816094|  VICENTE RAMOS|    4481|          -1|
|7519|2207002875190032|   JÚLIA CASTRO|    7519|          -1|
|1266|5979005712668585|VALENTINA CUNHA|    1266|          -1|
```

```
|8818|3039004288185454|  ANDRÉ PEREIRA|    8818|          -1|
|9176|8326004491766570|     DAVI LUCCA|    9176|          -1|
|4352|8175005043528487|  BRUNO REZENDE|    4352|          -1|
|7581|2276004575812746| MARIA FERNANDA|    7581|          -1|
|7158|3088004571584330|    YASMIN DIAS|    7158|          -1|
|6748|7121005067482756|      LUIGI PAZ|    6748|          -1|
|9350|4302006693504358|     DANIEL PAZ|    9350|          -1|
|1967|7614004019675322|    BRUNO PIRES|    1967|          -1|
|6312|3344004963126242|  ANTÔNIO ROCHA|    6312|          -1|
|4071|6180003840719159|   ENZO RIBEIRO|    4071|          -1|
|2674|8465002526749190| BRENDA RIBEIRO|    2674|          -1|
|2258|4298002822586857|   ESTHER SILVA|    2258|          -1|
|2827|6554006228271468|  YASMIN MORAES|    2827|          -1|
+----+----------------+---------------+--------+------------+


Qtd. registros tratados com '-1'(Conta): 0
20 registros aleatórios com tratamento de: -1 (Conta)
+---+----------+------------+--------+------------+
| id|num_cartao|nom_impresso|id_conta|id_associado|
+---+----------+------------+--------+------------+
+---+----------+------------+--------+------------+
```

Validar a integridade de relacionamento entre as tabelas Cartao e Movimento

In [30]:

```
df_nao_encontrado=spark.sql('''
    select tgt1.id_cartao

    from movimento_tgt tgt1

    left join cartao_tgt tgt2
    on tgt2.id=tgt1.id_cartao

    where tgt2.id is null
''')

qtd_nao_encontrado=df_nao_encontrado.count()

df_tratado=spark.sql('''
    select tgt1.*

    from movimento_tgt tgt1

    where tgt1.id_cartao=-1
    and tgt1.id<>-1
''')

qtd_tratados=df_tratado.count()

print("Validação da quantidade de registros: CSV x Target (Data Lake)")
print()
print(f"Qtd. registros não encontrados por problemas de integridade: {qtd_nao_encontrado}.")
print()
print(f"Qtd. registros tratados com '-1': {qtd_tratados}")
print("20 registros aleatórios com tratamento de: -1")
df_tratado.show()
```

```
[Stage 86:=========================================>             (149 + 1) / 200]
Validação da quantidade de registros: CSV x Target (Data Lake)

Qtd. registros não encontrados por problemas de integridade: 0.

Qtd. registros tratados com '-1': 0
20 registros aleatórios com tratamento de: -1
+---+-------------+-------------+--------------+---------+
| id|vlr_transacao|des_transacao|data_movimento|id_cartao|
+---+-------------+-------------+--------------+---------+
```

```
+---+------------+------------+-------------+---------+
```

Validar a integridade de relacionamento entre as tabelas Encerramento Conta e Conta

In [31]:

```
df_nao_encontrado=spark.sql('''
    select tgt1.id

    from encerramento_conta_tgt tgt1

    left join conta_tgt tgt2
    on tgt2.id=tgt1.id

    where tgt2.id is null
''')

qtd_nao_encontrado=df_nao_encontrado.count()

df_tratado=spark.sql('''
    select tgt1.*

    from encerramento_conta_tgt tgt1

    where tgt1.id=-1
    and tgt1.id<>-1
''')

qtd_tratados=df_tratado.count()

print("Validação da quantidade de registros: CSV x Target (Data Lake)")
print()
print(f"Qtd. registros não encontrados por problemas de integridade: {qtd_nao_encontrado}.")
print()
print(f"Qtd. registros tratados com '-1': {qtd_tratados}")
print("20 registros aleatórios com tratamento de: -1")
df_tratado.show()
```

```
Validação da quantidade de registros: CSV x Target (Data Lake)

Qtd. registros não encontrados por problemas de integridade: 0.

Qtd. registros tratados com '-1': 0
20 registros aleatórios com tratamento de: -1
+---+------------+-----------------+
| id|data_criacao|data_encerramento|
+---+------------+-----------------+
+---+------------+-----------------+
```

Validar a integridade de relacionamento entre as tabelas Fatura e Cartao

In [32]:

```
df_nao_encontrado=spark.sql('''
    select tgt1.id

    from fatura_tgt tgt1

    left join fatura_tgt tgt2
    on tgt2.id=tgt1.id

    where tgt2.id is null
''')

qtd_nao_encontrado=df_nao_encontrado.count()

df_tratado=spark.sql('''
```

```
    select tgt1.*

    from fatura_tgt tgt1

    where tgt1.id=-1
    and tgt1.id<>-1
''')

qtd_tratados=df_tratado.count()

print("Validação da quantidade de registros: CSV x Target (Data Lake)")
print()
print(f"Qtd. registros não encontrados por problemas de integridade: {qtd_nao_encontrado}.")
print()
print(f"Qtd. registros tratados com '-1': {qtd_tratados}")
print("20 registros aleatórios com tratamento de: -1")
df_tratado.show()
```

```
[Stage 99:=========================================>             (151 + 1) / 200]
Validação da quantidade de registros: CSV x Target (Data Lake)

Qtd. registros não encontrados por problemas de integridade: 0.

Qtd. registros tratados com '-1': 0
20 registros aleatórios com tratamento de: -1
+---+----------------------+----------+---------------------+--------------------+---------+
| id|data_vencimento_fatura|vlr_fatura|data_pagamento_fatura|qtd_dias_atraso_pgto|id_cartao|
+---+----------------------+----------+---------------------+--------------------+---------+
+---+----------------------+----------+---------------------+--------------------+---------+
```

Resultado dos testes

Teste unitário

Não foi identificado nenhuma irregulariedade em relação aos dados. Todos os dados de origem foram armazenados corretamente no Data Lake.

Teste integrado

Durante os testes, foi identificado irregularidades nos dados dos associados.

Durante os testes das informações de *conta* e *cartao*, foi identificado que existem 50 registros faltantes na tabela de associados.

O processo de carga tratou a carga corretamente preenchendo com o valor **-1** na ausência dos associados.

Simulação de Dados - Dados gerados para ocasionar o problema de integridade dos dados

As vezes podem ocorrer erros nos sistemas transacionais que geram inconsistências nos dados.

Essas inconsistências afetam diretamente os dados no Data Lake. Através dos processos de carga, podemos contornar alguns problemas de integridade relacional.

Para demonstrar melhor o processo de carga, foi excluído propositalmente 50 associados do arquivo de associados, deixando o ambiente mais próximo da realidade que encontramos nas organizações.

In [ ]: