ACADEMIA MULHERES EM TECH
DATA ENGINEER

PROJETO INTEGRADOR - GRUPO 5

accenture

GAMA ACADEMY

Spark Girls

Ana Caroline D'Oliveira

Caroline Ferraz

Laís Villa

Marina Varela

Paula Rocha

Raquel Reis

Raquel Zanatta

GAMA ACADEMY

accenture

Contextualização

Transactions_in
id
cliente_id
valor
data

Clients
id
nome
email
data_cadastro
telefone

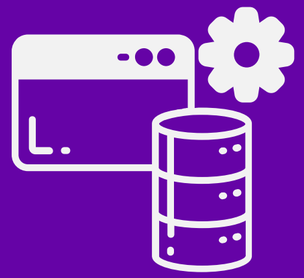
Transactions_out
id
cliente_id
valor
data

GAMA ACADEMY

accenture

Objetivos e Requisitos

#1
Script de migração em Python



#2
Modelo de Entidades e Relacionamentos

#3
Relatórios de análise em SQL e PowerBI

#4
Códigos versionados no GitHub

GAMA ACADEMY

accenture

Metodologias Organizacionais Utilizadas

Scrum

Trello

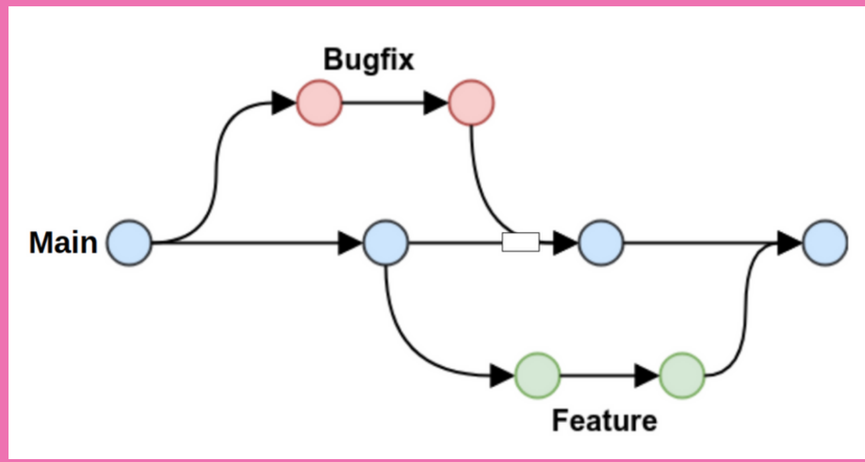
Zoom

Git Workflow

Metodologias Organizacionais Utilizadas

Versionamento no GitHub

print dos commits no github aqui



GAMA ACADEMY

accenture

Principais Ferramentas Utilizadas

Azure

Spark

Python

Power BI

GAMA ACADEMY

accenture

Script de migração utilizando Pyspark

Processo de ETL: Extract, Transform

- Importação dos dados
- Tratamentos dos dados
- Análise para identificar as fraudes

Transaction-out:

[5]:
```
transaction_out = spark.read.option("sep", ";").schema("id int, cliente_id int, valor double, data timestamp").option("header", "false").csv("/

transaction_out_sem_header = transaction_out.filter(~input_file_name().rlike("transaction-out-001.csv"))
transaction_out_com_header = spark.read.option("sep", ";")\
    .schema("id int, cliente_id int, valor double, data timestamp")\
    .option("header", "true")\
    .csv("/home/azureuser/transaction-out/transaction-out-001.csv")

transaction_out = transaction_out_com_header.union(transaction_out_sem_header)

transaction_out.count()
transaction_out.show()
```

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

Análises/Transformações para identificar as fraudes:

Unindo as duas tabelas de transações (in, out):

[6]:
```
# df_transactions ----- O Dataframe de todas as transações (in e out)
df_transactions = (transaction_in.union(transaction_out))

print(f"Contagem de todas as transações: {df_transactions.count()}")

df_transactions.show(truncate=False)
```

```
Contagem de todas as transações: 7272
+----+----------+-----+-------------------+
|id  |cliente_id|valor|data               |
+----+----------+-----+-------------------+
|8615|586       |0.2  |2022-01-19 20:12:26|
|8613|586       |0.2  |2022-01-19 20:11:25|
|8611|586       |0.2  |2022-01-19 20:10:05|
|8606|910       |300.0|2022-01-19 19:59:36|
|8604|76        |100.0|2022-01-18 12:48:14|
```

GAMA ACADEMY

accenture

Script de migração utilizando Pyspark

Processo de ETL: Extract, Transform

- Análise para identificar as fraudes
- Transformação dos dados gerando um novo DataFrame

Adicionando na tabela df_transaction uma coluna para a diferença de tempo entre as transações:

7]:
```
from pyspark.sql.functions import lag, datediff, col, when, lit,desc, asc
from pyspark.sql.window import Window

# Essa window irá particionar a tabela através do cliente_id e ordenar essa partição por data
window_spec = Window.partitionBy("cliente_id").orderBy("data")

# Essa parte cria a coluna "diferenca_de_tempo" para o tempo entre as transações (obs.: O tempo será 0, quando a diferenca de tempo for nula, c
diferenca_de_tempo = (col("data").cast("long") -
                      lag(col("data")).over(window_spec).cast("long"))

df_transactions = df_transactions.withColumn("diferenca_de_tempo", when(diferenca_de_tempo.isNull(), lit(0)).otherwise(diferenca_de_tempo))

df_transactions.show()
```

```
+---+----------+-----+-------------------+------------------+
| id|cliente_id|valor|               data|diferenca_de_tempo|
+---+----------+-----+-------------------+------------------+
```

Adicionando a coluna 'fraudado', do tipo boolean, à tabela df_transactions:

[ ]:
```
from pyspark.sql.functions import col, when
```

[ ]:
```
df_transactions = df_transactions\
    .withColumn("fraudado", when((col("diferenca_de_tempo") <= 120) & (col("diferenca_de_tempo") != 0), True)\
    .otherwise(False))
```

Analisando o DataFrame gerado:

[ ]:
```
df_transactions.printSchema()
```

[ ]:
```
df_transactions.count()
```

GAMA ACADEMY

accenture

Script de migração utilizando Pyspark

Processo de ETL: Load

- Inserção dos dados transformados no banco de dados SQL Server na nuvem Azure

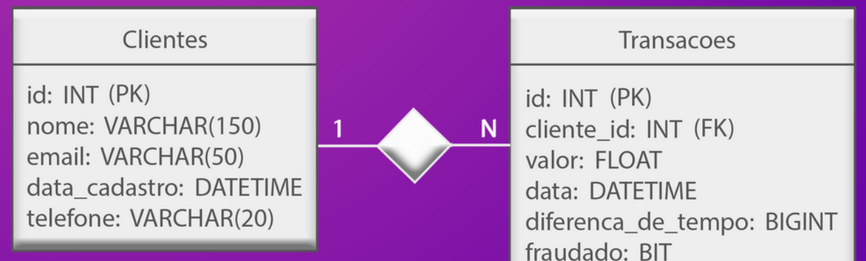
Migração dos dados do Spark para o banco SQL Server (tabela transações)

[ ]:

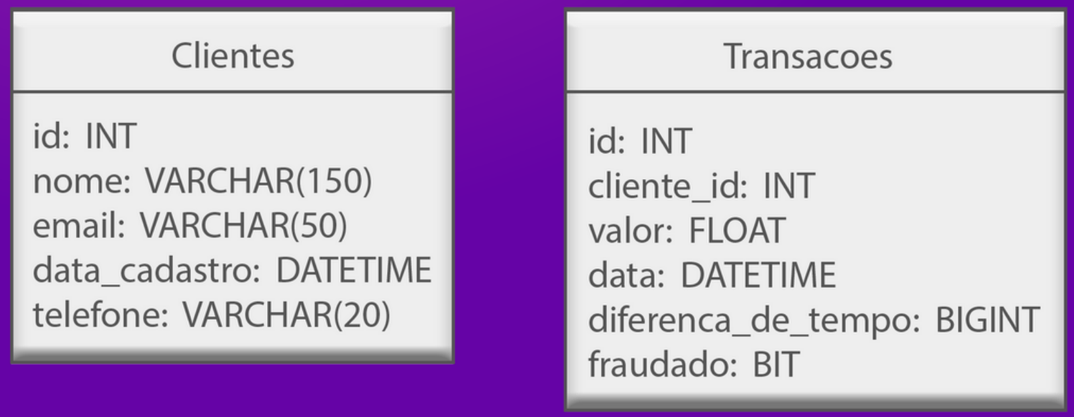
```
import pyodbc
from pyspark.sql import SparkSession
from pyspark.sql.functions import col

# Criar uma sessão do Spark
spark = SparkSession.builder.appName("Exemplo").getOrCreate()

# Criando uma conexão com o SQL Server

conn = pyodbc.connect("Driver={ODBC Driver 18 for SQL Server};Server=tcp:projeto-integrador5.database.windows.net,1433;Database=projeto_integra
cursor = conn.cursor()


# Cria a tabela "transacoes" caso ela não exista ou exclui a tabela existente e cria uma nova
cursor.execute("IF OBJECT_ID('transacoes', 'U') IS NOT NULL "
               "DROP TABLE transacoes; "
               "CREATE TABLE transacoes "
               "( "
               "id INT PRIMARY KEY, "
               "cliente_id INT, "
               "valor FLOAT, "
               "data DATETIME, "
               "diferenca_de_tempo INT, "
               "fraudado BIT "
               "); ")
conn.commit()
```

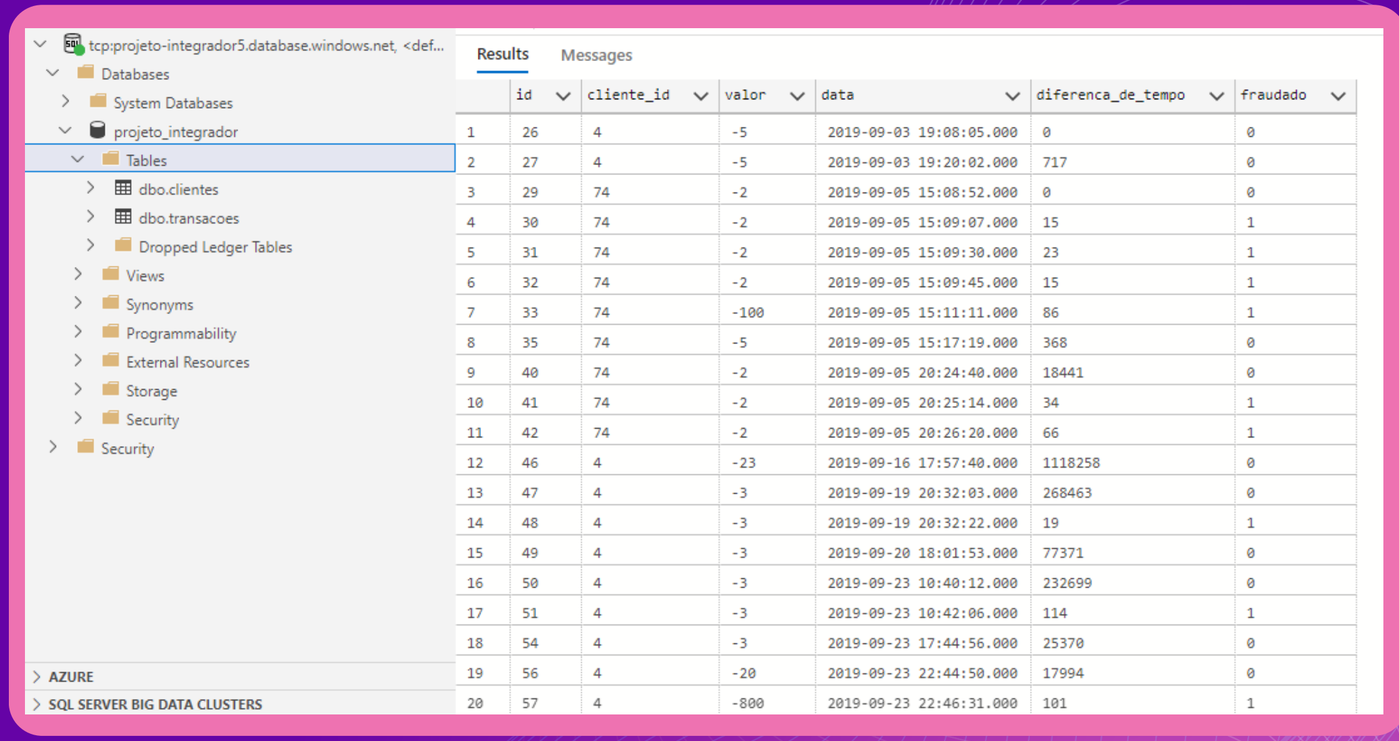
```
array_transaction = [row.asDict() for row in df_transactions.collect()]


for transacao in array_transaction:
    cursor.execute("INSERT INTO transacoes (id, cliente_id, valor, data, diferenca_de_tempo, fraudado) "
            "VALUES (?, ?, ?, ?, ?, ?)",
            transacao["id"], transacao["cliente_id"], transacao["valor"], transacao["data"], transacao["diferenca_de_tempo"], transacao["frauda
    conn.commit()

cursor.close()
conn.close()
```

Modelo relacional gerado



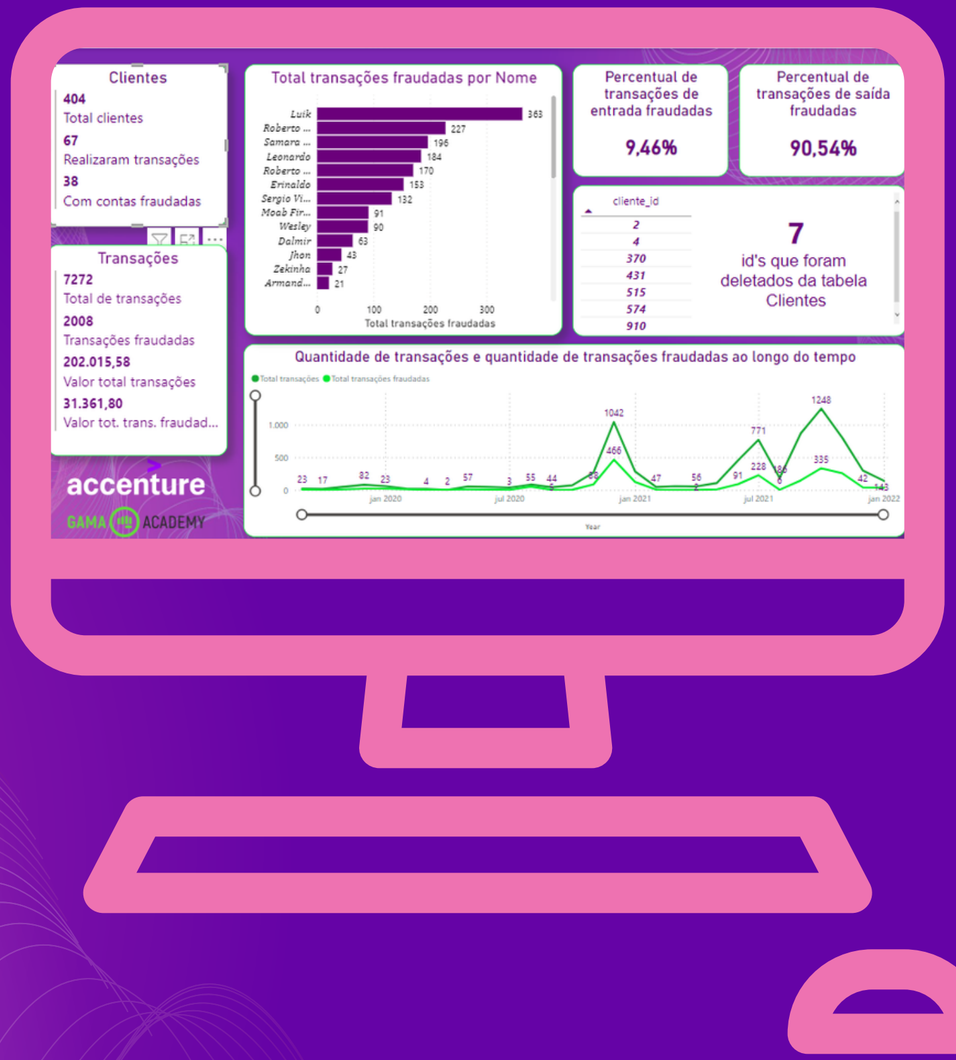
X





	id	cliente_id	valor	data	diferenca_de_tempo	fraudado
1	26	4	-5	2019-09-03 19:08:05.000	0	0
2	27	4	-5	2019-09-03 19:20:02.000	717	0
3	29	74	-2	2019-09-05 15:08:52.000	0	0
4	30	74	-2	2019-09-05 15:09:07.000	15	1
5	31	74	-2	2019-09-05 15:09:30.000	23	1
6	32	74	-2	2019-09-05 15:09:45.000	15	1
7	33	74	-100	2019-09-05 15:11:11.000	86	1
8	35	74	-5	2019-09-05 15:17:19.000	368	0
9	40	74	-2	2019-09-05 20:24:40.000	18441	0
10	41	74	-2	2019-09-05 20:25:14.000	34	1
11	42	74	-2	2019-09-05 20:26:20.000	66	1
12	46	4	-23	2019-09-16 17:57:40.000	1118258	0
13	47	4	-3	2019-09-19 20:32:03.000	268463	0
14	48	4	-3	2019-09-19 20:32:22.000	19	1
15	49	4	-3	2019-09-20 18:01:53.000	77371	0
16	50	4	-3	2019-09-23 10:40:12.000	232699	0
17	51	4	-3	2019-09-23 10:42:06.000	114	1
18	54	4	-3	2019-09-23 17:44:56.000	25370	0
19	56	4	-20	2019-09-23 22:44:50.000	17994	0
20	57	4	-800	2019-09-23 22:46:31.000	101	1

Relatórios de Análise



Conclusão

GAMA ACADEMY

accenture

AGRADECEMOS PELA DISPONIBILIDADE!

Ana Caroline D'Oliveira

Caroline Ferraz

Laís Villa

Marina Varela

Paula Rocha

Raquel Reis

Raquel Zanatta

accenture

GAMA ACADEMY