

Engenharia de Dados

Cessão de Crédito Consignado

Autor: Alexander Almeida

Última atualização: quarta-feira, 3 de julho de 2024

ÍNDICE

| | |
|--|----------|
| 1. MVP | 3 |
| 1.1. Objetivo | 3 |
| 1.2. Análise | 3 |
| 1.3. Metadado | 3 |
| 1.4. Arquitetura do ETL | 4 |
| 1.4.1. Camada Bronze (Extração) | 5 |
| 1.4.2. Camada Prata (Transformação) | 6 |
| 1.4.3. Camada Ouro (Carga) | 8 |
| 1.4.3.1. Dataproc | 8 |
| 1.4.3.2. Linhagem | 10 |
| 1.4.3.3. Dataplex (Data Quality) | 10 |
| 1.4.4. Looker Stúdio (Análises) | 11 |
| 1.4.4.1. Volume de idade por UF | 12 |
| 1.4.4.2. Volume da margem de saldo por UF | 14 |
| 1.4.4.3. Volume da margem de saldo por UF e grupo de idade | 16 |
| 1.4.5. Conclusão | 17 |
| 1.4.5.1. Considerações | 17 |
| 1.4.5.2. Recomendações | 17 |

1. MVP

1.1. Objetivo

O objetivo principal é analisar os dados para ser mais assertivo na venda do crédito consignado. A regra principal consiste em ser menor que 85 anos de idade, com margem de saldo superior à 35% dos parcelamentos vigentes.


O dataset possui 6176 instâncias e 9 atributos, com um mix de tipos de dados categóricos e numéricos.

1.2. Análise

- Qual o volume de Idade por UF?
- Qual o volume da Margem de Saldo por UF?
- Qual a relação entre UF, Idade e Margem de Saldo?

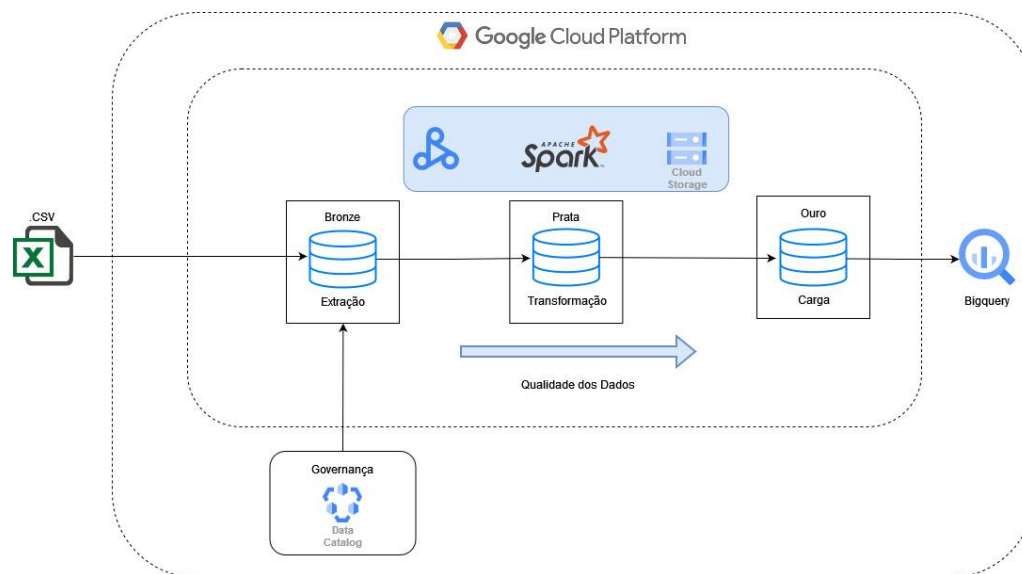
Com estas definições, poderemos montar ações estratégicas para cessão de crédito

1.3. Metadado

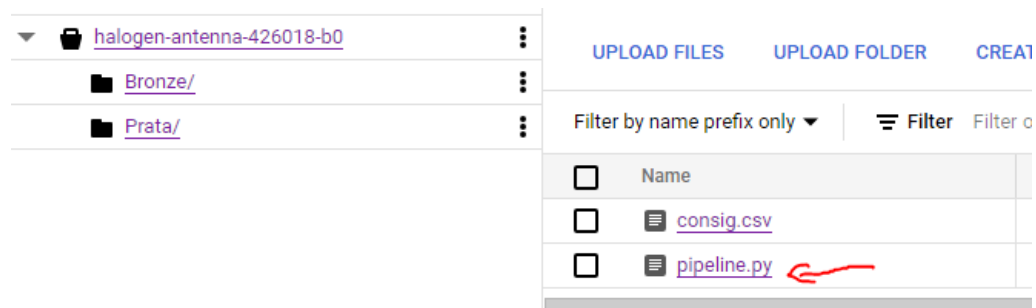
| <input type="checkbox"/> | Field name | Type | Mode | Key | Collation | Default value | Policy tags  | Description |
|--------------------------|--------------------|---------|----------|-----|-----------|---------------|---|---|
| <input type="checkbox"/> | cpf | STRING | NULLABLE | - | - | - | - | CPF do beneficiário |
| <input type="checkbox"/> | nome | STRING | NULLABLE | - | - | - | - | Nome do beneficiário |
| <input type="checkbox"/> | sexo | STRING | NULLABLE | - | - | - | - | Sexo do beneficiário |
| <input type="checkbox"/> | idade | INTEGER | NULLABLE | - | - | - | - | Idade do beneficiário |
| <input type="checkbox"/> | uf | STRING | NULLABLE | - | - | - | - | Unidade Federativa de origem do beneficiário |
| <input type="checkbox"/> | situacao_funcional | STRING | NULLABLE | - | - | - | - | Situação funcional: Aposentado/Pensionista |
| <input type="checkbox"/> | tipo_contrato | STRING | NULLABLE | - | - | - | - | Estável: Aposentadoria Estável / Não Estável: Pode perder a aposentadoria |
| <input type="checkbox"/> | pmt_valor | FLOAT | NULLABLE | - | - | - | - | Valor total dos parcelamentos vigentes |
| <input type="checkbox"/> | margem_saldo | FLOAT | NULLABLE | - | - | - | - | Valor total disponível para cessão de crédito |
| <input type="checkbox"/> | conceder | INTEGER | NULLABLE | - | - | - | - | Regra do negócio para cessão de crédito, onde 0=CONCEDER e 1=Não conceder |

1.4. Arquitetura do ETL

Estaremos utilizando a arquitetura medalhão neste MVP, conforme figura abaixo:



Toda construção é realizada através de uma pipeline em pyspark (pipeline.py)



1.4.1. Camada Bronze (Extração)

Na camada bronze carreguei o arquivo .CSV no Cloud Storage GCP, pasta Bronze, em seu estado bruto. A base utilizada é particular e real, mas para apresentação neste MVP, foi necessária a anonimização do campo CPF. Esse processo de extração não foi automatizado, devido a limitação da conta "free" na GCP. Desta forma, realizei o UPLOAD manualmente, conforme figura abaixo:

```
[ ]: """
*****
Devido a limitação na nuvem GCP, devido ao plano Free, não conseguirei demonstrar o upload do arquivo pela pipeline. Desta forma, realizarei o UPLOAD manualmente no bucket
*****

# Definição do projeto na GCP
project_id = "halogen-antenna-426018-b0"
bucket_name = "halogen-antenna-426018-b0"
destination_blob_name = "Prata"

github_file_url = "https://raw.githubusercontent.com/AlexanderAlmeida/pos_graduacao/master/consig.csv"

def download_and_upload_to_gcs(github_file_url, bucket_name, destination_blob_name, project_id):
    # Download .CSV do GitHub
    response = requests.get(github_file_url)
    if response.status_code == 200:
        csv_data = response.content
    else:
        raise Exception(f"Error downloading file: {response.status_code}")

    # Criação do arquivo temporário para armazenar os dados do download
    with tempfile.NamedTemporaryFile(delete=False) as temp_file:
        temp_file.write(csv_data)
        filename = temp_file.name # Get the temporary file path

    # Autenticar no Cloud Storage e criar no bucket
    storage_client = storage.Client(project=project_id)
    bucket = storage_client.get_bucket(bucket_name)

    # Upload do .CSV para o Cloud Storage
    blob = bucket.blob(destination_blob_name)
    blob.upload_from_filename(filename)

    # Delete the temporary file after upload
    # (Optional, the file will be automatically deleted when the program exits)
    os.remove(filename)

download_and_upload_to_gcs(github_file_url, bucket_name, destination_blob_name, project_id)

print(f"File downloaded from GitHub and uploaded to bucket: {bucket_name}/{destination_blob_name}")
"""
```

Folder browser

▼

halogen-antenna-426018-b0

⋮

▶

bronze/

⋮

▶

Prata/

⋮

Buckets > halogen-antenna-426018-b0 > Bronze

UPLOAD FILES

UPLOAD FOLDER

CREATE FOLDER

TRANSFER DATA

MANAGE HOLDS

EDIT RETENTION



DOWNLOAD

DELETE

Filter by name prefix only

Filter

Filter objects and folders

| <input type="checkbox"/> | Name | Size | Type | Created | Storage class | Last modified |
|--------------------------|--|----------|--------------------------|-----------------------|---------------|-----------------------|
| <input type="checkbox"/> |  consig.csv | 609.3 KB | text/csv | 23 Jun 2024, 12:34:17 | Standard | 23 Jun 2024, 12:34:17 |
| <input type="checkbox"/> |  pipeline.py | 6.2 KB | application/octet-stream | 1 Jul 2024, 12:41:11 | Standard | 1 Jul 2024, 12:41:11 |

Referências:

pipeline.py:

https://raw.githubusercontent.com/AlexanderAlmeida/pos_graduacao/master/pipeline.py

Consig.csv:

https://raw.githubusercontent.com/AlexanderAlmeida/pos_graduacao/master/consig.csv

1.4.2. Camada Prata (Transformação)

Na camada prata carreguei o dataframe com spark para início dos tratamentos de dados. Criei um contador de instâncias para observabilidade.

```
In [3]: #####
# CANADA BRONZE (Ingestão)
#####

# Leia o dataset do Cloud Storage
df_fe = spark.read.csv("gs://halogen-antenna-426018-b0/Bronze/consig.csv", header=True, inferSchema=True)
df_fe.show()

# Contando o número de linhas do dataset para observabilidade
num_rows_csv = df_fe.count()
print(f"Number of rows in CSV: {num_rows_csv}")
```

| | | | | | | | | |
|------------|----------------------|-----------|----|----|---------------------|---------|--------------------|--------------------|
| 0000**2074 | CRISTIANO SANTOS ... | MASCULINO | 37 | RS | ATIVO PERMANENTE | ESTAVEL | 45.07 | 1361.11 |
| 0000**3748 | MARIA DA GLORIA V... | FEMININO | 52 | RJ | ATIVO PERMANENTE | ESTAVEL | 37.15 | 2192.17 |
| 0000**3204 | CEREJA KAZUKO NAK... | FEMININO | 80 | RS | APOSENTADO | ESTAVEL | 1922.6 | 5160.92 |
| 0000**3155 | SILBERTO DOS SANT... | MASCULINO | 36 | MG | ATIVO PERMANENTE | ESTAVEL | 748.4200000000001 | 2064.08 |
| 0000**1234 | IRIS PEDRO DE OLI... | FEMININO | 75 | GO | APOSENTADO | ESTAVEL | 9723.39 | 19903.32 |
| 0000**7742 | JANICE CAMPOS MOTTA | FEMININO | 55 | RJ | APOSENTADO | ESTAVEL | 576.3299999999999 | 4161.24 |
| 0000**4706 | FABIO DE JESUS RI... | MASCULINO | 52 | RJ | ATIVO PERMANENTE | ESTAVEL | 550.29 | 3251.9 |
| 0000**6725 | GIOVANI AZEVEDO S... | MASCULINO | 56 | SC | NAO INFORMADO | ESTAVEL | 119.0 | 598.8 |
| 0000**2709 | JERONIMO DOS SANT... | MASCULINO | 58 | RJ | ATIVO PERMANENTE | ESTAVEL | 1862.14 | 2880.3500000000004 |
| 0000**0500 | MARCELLO PORTELA ... | MASCULINO | 40 | MS | ATIVO PERMANENTE | ESTAVEL | 831.53 | 11245.95 |
| 0000**4353 | JOSE WEVERGTHON A... | MASCULINO | 82 | SP | APOSENTADO | ESTAVEL | 683.21 | 15209.95 |
| 0000**0065 | VANDERLEI DOMINGU... | MASCULINO | 45 | GO | CELETISTA EMPREGADO | ESTAVEL | 1640.6699999999998 | 2337.4800000000005 |
| 0000**4729 | JOSILEI TRINDADE ... | FEMININO | 50 | RJ | CEDIDO SUS LEI 8270 | ESTAVEL | 302.18 | 2010.96 |
| 0000**4287 | JOANA CORREA DE S... | FEMININO | 82 | PA | APOSENTADO | ESTAVEL | 45.99 | 1384.39 |
| 0000**3127 | LARISSA DE SOUZA ... | FEMININO | 36 | DF | CELETISTA EMPREGADO | ESTAVEL | 3887.82 | 8048.039999999999 |
| 0000**9582 | MARIA EUNICE DE J... | FEMININO | 76 | BA | APOSENTADO | ESTAVEL | 50.06 | 713.1 |

-----+
only showing top 20 rows
Number of rows in CSV: 6176

Em seguida, armazenei o dataset no BigQuery, eu seu estado ainda bruto, para qualidade dos dados e linhagem.

```
In [4]: # Autenticação no BigQuery
cliente_bq = bigquery.Client()

# Define o nome do projeto e do conjunto de dados no BigQuery
dataset_id = "CREDITO_CONSIGNADO"

# Armazenando os dados brutos no bigquery
df_fe.write.format("bigquery").option("temporaryGcsBucket", "").option("writeMethod", "DIRECT").option("project", project_id).opt

# Verificando a quantidade de registros importados para observabilidade

query = f"SELECT COUNT(*) as total_rows FROM `halogen-antenna-426018-b0.CREDITO_CONSIGNADO.tbl_consignado_raw`"
query_job = cliente_bq.query(query)
result = query_job.result()
num_rows_bq = list(result)[0].total_rows
print(f"Number of rows in BigQuery table: {num_rows_bq}")
```

Number of rows in BigQuery table: 6176

| tbl_consignado_raw | | | | | | | | | |
|--------------------|------------|----------------------------|-----------|-------|----|---------------------|---------------|-------------|-------------|
| Row | cpf | nome | sexo | idade | uf | situacao_funcional | tipo_contrato | pmt_valor | margem_sal |
| 1 | 0000**1287 | EDILSON NASCIMENTO DA SIL. | MASCULINO | 80 | PA | APOSENTADO | ESTAVEL | 72.02 | 2498.8 |
| 2 | 0000**2760 | LARA TANIA GONCALVES | FEMININO | 59 | RJ | ATIVO PERMANENTE | ESTAVEL | 946.110000. | 10053.2 |
| 3 | 0000**7125 | SUZANA MENEZES GARCIA | FEMININO | 38 | MS | ATIVO PERMANENTE | ESTAVEL | 1250.04 | 3126.4 |
| 4 | 0000**2468 | CARLOS ROBERTO RIBEIRO DE. | FEMININO | 80 | PE | APOSENTADO | ESTAVEL | 721.14 | 20262.9 |
| 5 | 0000**2074 | CRISTIANO SANTOS ROSSONI | MASCULINO | 37 | RS | ATIVO PERMANENTE | ESTAVEL | 45.07 | 1361.1 |
| 6 | 0000**3748 | MARIA DA GLORIA VONCELA. | FEMININO | 52 | RJ | ATIVO PERMANENTE | ESTAVEL | 37.15 | 2302.1 |
| 7 | 0000**1204 | CEREJA KAZUO NAKAUCHI | FEMININO | 80 | RS | APOSENTADO | ESTAVEL | 1922.6 | 5140.9 |
| 8 | 0000**9155 | SILBERTO DOS SANTOS SILVA | MASCULINO | 36 | MS | ATIVO PERMANENTE | ESTAVEL | 748.420000. | 2064.0 |
| 9 | 0000**1234 | IRIS PEDRO DE OLIVEIRA | FEMININO | 75 | GO | APOSENTADO | ESTAVEL | 9723.39 | 19903.3 |
| 10 | 0000**7742 | JANICE CAMPOS MOTTA | FEMININO | 55 | RJ | APOSENTADO | ESTAVEL | 576.329999. | 4161.2 |
| 11 | 0000**4706 | FABIO DE JESUS RIBEIRO | MASCULINO | 52 | RJ | ATIVO PERMANENTE | ESTAVEL | 550.29 | 3251. |
| 12 | 0000**6725 | GIOVANI AZEVEDO SAO LOTO | MASCULINO | 56 | SC | NAO INFORMADO | ESTAVEL | 119.0 | 598. |
| 13 | 0000**2709 | JEONIMO DOS SANTOS BAR. | MASCULINO | 58 | RJ | ATIVO PERMANENTE | ESTAVEL | 1862.14 | 2880.30000. |
| 14 | 0000**6500 | MARCELLO PORTELA SILVA | MASCULINO | 40 | MS | ATIVO PERMANENTE | ESTAVEL | 891.58 | 11245.9 |
| 15 | 0000**4303 | JOSE REVERTHON AGUIAR. | MASCULINO | 82 | SP | APOSENTADO | ESTAVEL | 483.21 | 15209.9 |
| 16 | 0000**0505 | VANDELLEI DOMINGUES FAGU. | MASCULINO | 45 | GO | CELETISTA EMPREGADO | ESTAVEL | 1640.66999. | 2337.48000. |
| 17 | 0000**4729 | JOSILRI TRINHADE RANGEL | FEMININO | 50 | RJ | CEDIDO SUS LI 8270 | ESTAVEL | 302.18 | 2010.9 |

Na sequência foi observado no dataset, CPFs duplicados, devido a anonimização. Dito isto, foram removidas as duplicidades.

```
In [5]: #####
#CAMADA PRATA (Transformação)
#####

## Mantenha apenas a primeira ocorrência por CPF
# Verificando duplicidade no CPF, devido a anonimização do campo para este MVP
dataset_sem_duplicatas = df_fe.withColumn("row_num", F.row_number().over(Window.partitionBy("CPF").orderBy(F.col("cpf").desc())))
dataset_sem_duplicatas = dataset_sem_duplicatas.filter(F.col("row_num") == 1)

dataset_sem_duplicatas.show()

dataset_sem_duplicatas = dataset_sem_duplicatas.filter(F.col("row_num") == 1).drop("row_num")
df_fe = dataset_sem_duplicatas
```

Como a análise será realizada por faixa de idade, foi necessária adição das mesmas ao dataset. Ao término, armazenei o novo dataset em formato parquet ao Cloud Storage GCP, na pasta Prata.

```
# Crie colunas para cada faixa etária
df_fe = df_fe.withColumn("Faixa_10_20", F.when(df_fe["idade"].between(10, 20), 1).otherwise(0))
df_fe = df_fe.withColumn("Faixa_21_30", F.when(df_fe["idade"].between(21, 30), 1).otherwise(0))
df_fe = df_fe.withColumn("Faixa_31_40", F.when(df_fe["idade"].between(31, 40), 1).otherwise(0))
df_fe = df_fe.withColumn("Faixa_41_50", F.when(df_fe["idade"].between(41, 50), 1).otherwise(0))
df_fe = df_fe.withColumn("Faixa_51_60", F.when(df_fe["idade"].between(51, 60), 1).otherwise(0))
df_fe = df_fe.withColumn("Faixa_61_70", F.when(df_fe["idade"].between(61, 70), 1).otherwise(0))
df_fe = df_fe.withColumn("Faixa_71_80", F.when(df_fe["idade"].between(71, 80), 1).otherwise(0))
df_fe = df_fe.withColumn("Faixa_81_90", F.when(df_fe["idade"].between(81, 90), 1).otherwise(0))

df_fe.show()
```

| cpf | nome | sexo | idade | uf | situacao_funcional | tipo_contrato | pmt_valor | margem_saldo |
|-------------|-----------------------|-----------|-------|----|---------------------|---------------|--------------------|---------------------|
| 0**0***3287 | CARLOS ALBERTO S ... | FEMININO | 76 | AM | APOSENTADO | ESTAVEL | 1540.4199999999998 | 4998.18 |
| 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 0**2***7164 | ARIANA MARIA DE C ... | FEMININO | 37 | MT | CELETISTA EMPREGADO | ESTAVEL | 3508.71 | 3184.28 |
| 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 0**4***3744 | ELIANE GOMES DA S ... | FEMININO | 55 | RJ | APOSENTADO | ESTAVEL | 2948.42 | -3172.8000000000006 |
| 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 0**6***0365 | DANNIEL ROCHA DO ... | MASCULINO | 38 | PI | ATIVO PERMANENTE | ESTAVEL | 4678.62 | 10253.44 |
| 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 0**8***4172 | SUZANA DE MATTOS ... | FEMININO | 84 | MG | BENEFICIARIO PENSAO | NAO ESTAVEL | 1584.85 | 381.75 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 0**9***0463 | GILSON EDMAR GONC... | FEMININO | 78 | PE | APOSENTADO | ESTAVEL | 652.04 | 8514.0 |

1.4.3. Camada Ouro (Carga)

Basicamente o dataset foi agrupado com as colunas UF, Margem_Saldo, Grupo de faixas e adicionado o campo id (identity). Após isso, gravamos o dataset no bigquery como tabela fato (FLAT), para realização das análises.

```
#####
#CARGADA OURO (Carga)
#####

# Calcule a contagem por faixa etária e UF
df_fe = df_fe.groupBy(["uf"]).agg(F.sum(F.col("margem_saldo")).alias("Margem_Saldo"), F.sum("Faixa_10_20").alias("Faixa_10_20"), F.sum("Faixa_21_30").alias("Faixa_21_30"), F.sum("Faixa_31_40").alias("Faixa_31_40"), F.sum("Faixa_41_50").alias("Faixa_41_50"), F.sum("Faixa_51_60").alias("Faixa_51_60"), F.sum("Faixa_61_70").alias("Faixa_61_70"), F.sum("Faixa_71_80").alias("Faixa_71_80"), F.sum("Faixa_81_90").alias("Faixa_81_90"))

# Criação do campo ID
df_fe = df_fe.withColumn('id', F.monotonically_increasing_id())

# Reordenando as colunas e deletando o último campo
df_fe = df_fe.select('id', *df_fe.columns[1:])

# Exibindo o dataset completo
df_fe.show()

# Salve o dataset processado no BigQuery
df_fe.write.format("bigquery").option("temporaryGcsBucket", "").option("writeMethod", "DIRECT").option("project", project_id).option("dataset", dataset_id).option("table", "tbl_fato")

# Parar a sessão SparkSession
spark.stop()
```

| id | uf | Margem_Saldo | Faixa_10_20 | Faixa_21_30 | Faixa_31_40 | Faixa_41_50 | Faixa_51_60 | Faixa_61_70 | Faixa_71_80 | Faixa_81_90 |
|----|----|--------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| 0 | MS | 253197.91000000003 | 1 | 1 | 84 | 5 | 4 | 5 | 3 | 4 |
| 1 | CE | 1248672.97 | 1 | 0 | 105 | 7 | 3 | 6 | 73 | 93 |
| 2 | MG | 1888138.9199999974 | 0 | 1 | 21 | 167 | 34 | 32 | 62 | 110 |
| 3 | DF | 3392125.6299999985 | 0 | 8 | 417 | 47 | 20 | 6 | 82 | 108 |
| 4 | RO | 40581.68 | 1 | 13 | 21 | 1 | 1 | 2 | 4 | 0 |
| 5 | AM | 610785.1200000005 | 1 | 13 | 37 | 6 | 0 | 0 | 42 | 34 |
| 6 | MT | 375235.17000000016 | 0 | 0 | 55 | 4 | 4 | 4 | 10 | 13 |
| 7 | SP | 1064686.4700000007 | 1 | 2 | 48 | 28 | 18 | 142 | 78 | 30 |
| 8 | PB | 315985.21 | 0 | 1 | 9 | 59 | 9 | 6 | 14 | 9 |
| 9 | BA | 863103.3799999999 | 0 | 0 | 67 | 58 | 10 | 10 | 38 | 52 |
| 10 | SE | 280744.18000000005 | 0 | 0 | 35 | 19 | 1 | 2 | 6 | 1 |
| 11 | RJ | 1361111.9599999997 | 0 | 1 | 15 | 57 | 681 | 130 | 120 | 64 |
| 12 | AC | 64217.08999999999 | 0 | 8 | 11 | 1 | 2 | 1 | 5 | 3 |
| 13 | PR | 581894.5800000003 | 0 | 1 | 26 | 17 | 3 | 7 | 34 | 42 |
| 14 | AP | 116248.91000000002 | 0 | 6 | 22 | 0 | 1 | 0 | 18 | 6 |
| 15 | RR | 16245.420000000002 | 1 | 8 | 19 | 4 | 1 | 1 | 1 | 0 |
| 16 | TO | 154264.0 | 0 | 1 | 36 | 8 | 4 | 0 | 2 | 1 |
| 17 | ES | 215858.67 | 0 | 0 | 1 | 8 | 27 | 5 | 6 | 5 |
| 18 | ND | 1498.5 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 19 | AL | 124588.56 | 0 | 0 | 5 | 16 | 4 | 5 | 1 | 8 |

only showing top 20 rows

halogen-antenna-426018-b0

Queries

Notebooks

Data canvases

External connections

CREDITO_CONSIGNADO

tbtl_consignado_raw

tbtl_fato

consignado_raw

Filter

Enter property name or value

| <input type="checkbox"/> | Field name | Type | Mode | Key | Collation | Default value | Policy tags | Description |
|--------------------------|--------------|---------|----------|-----|-----------|---------------|-------------|--|
| <input type="checkbox"/> | id | INTEGER | REQUIRED | - | - | - | - | campo de identificação único |
| <input type="checkbox"/> | uf | STRING | NULLABLE | - | - | - | - | unidade federativa |
| <input type="checkbox"/> | Margem_Saldo | FLOAT | NULLABLE | - | - | - | - | Valor disponível para cessão de crédito |
| <input type="checkbox"/> | Faixa_10_20 | INTEGER | NULLABLE | - | - | - | - | Faixa etária agrupada entre 10 e 20 anos |
| <input type="checkbox"/> | Faixa_21_30 | INTEGER | NULLABLE | - | - | - | - | Faixa etária agrupada entre 21 e 30 anos |
| <input type="checkbox"/> | Faixa_31_40 | INTEGER | NULLABLE | - | - | - | - | Faixa etária agrupada entre 31 e 40 anos |
| <input type="checkbox"/> | Faixa_41_50 | INTEGER | NULLABLE | - | - | - | - | Faixa etária agrupada entre 41 e 50 anos |
| <input type="checkbox"/> | Faixa_51_60 | INTEGER | NULLABLE | - | - | - | - | Faixa etária agrupada entre 51 e 60 anos |
| <input type="checkbox"/> | Faixa_61_70 | INTEGER | NULLABLE | - | - | - | - | Faixa etária agrupada entre 71 e 80 anos |
| <input type="checkbox"/> | Faixa_71_80 | INTEGER | NULLABLE | - | - | - | - | Faixa etária agrupada entre 81 e 90 anos |
| <input type="checkbox"/> | Faixa_81_90 | INTEGER | NULLABLE | - | - | - | - | - |

OBS.: Utilizamos o **modelo flat** para o data warehouse.

1.4.3.1. Dataproc

Selecionamos o Google DataProc como gerenciador do Spark e Hadoop, para execução de toda pipeline deste projeto. Criamos o JOB denominado Pipeline_FULLL, apontando para o arquivo pyspark (pipeline.py), armazenado no bucket Bronze. Abaixo a saída de todo fluxo:


```
Job ID Pipeline_FULL
Job UUID 01bccf7e-c7bf-4d7b-9264-680c8b842f9
Type Dataproc job
Status Succeeded
```

Output LINQ WRAP: OFF

Spark jobs take ~60 seconds to initialise resources.

```
24/07/01 18:02:28 INFO SparkEnv: Registering RemoteOutputTracker
24/07/01 18:02:28 INFO SparkEnv: Registering BlockManagerMaster
24/07/01 18:02:28 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
24/07/01 18:02:28 INFO SparkEnv: Registering DataCommitCoordinator
24/07/01 18:02:28 INFO DefaultConfiguration: Connecting to ResourceManager at cluster-f371-m-us-central-f-c.halgain-antenna-4260818-eu.internal:12138.o.318032
24/07/01 18:02:28 INFO AMSPProvider: Connecting to Application History server at cluster-f371-m-us-central-f-c.halgain-antenna-4260818-eu.internal:12138.o.318020
24/07/01 18:02:28 INFO Configuration: resource-types.xml not found
24/07/01 18:02:28 INFO ResourceUtils: Unable to find "resource-types.xml".
24/07/01 18:02:30 INFO VarClientMetrics: Submitted application resourceTypes_1739490482878f_0084
24/07/01 18:02:31 INFO DefaultConfiguration: Connecting to ResourceManager at cluster-f371-m-us-central-f-c.halgain-antenna-4260818-eu.internal:12138.o.318030
24/07/01 18:02:33 INFO MetricsConfig: Loaded properties from hadoop-metrics.properties
24/07/01 18:02:33 INFO MetricsRegistryImpl: Scheduled metric snapshot period at 10 second(s).
24/07/01 18:02:33 INFO MetricsRegistryImpl: google-hadoop-file-system metrics system started
24/07/01 18:02:34 INFO GoogleCloudStorageImpl: Ignoring exception of type GoogleCloudStorageException; verified object already exists with desired state.
gfs://.../dataproc-temp-us-central-1-54673246904-a1ef6ff91fa65966-d6ac-4568-802-0905464...
=====
(0000)=====0628|ECCOLIUN NASCHICHEN |PASCULLI|      88| PA |    APPENDASOAT |ESTABLISHED|       72.82 |        2456.88| 0|
(0000)=====2740|LARA TIANA OCHOAVALVES |FENNIDNO|     59| KJ |    ACTIVE PERM|ESTABLISHED|   19.461.100.000.000 |         10853.26| 0|
(0000)=====7725|SUZANA FERNES DA G...|FENNIDNO|     38| FS |    ACTIVE PERM|ESTABLISHED|   1259.04 |        3126.00| 0|
(0000)=====2468|CARLOS ROBERTO RE ...|FENNIDNO|     88| PE |    APPENDASOAT |ESTABLISHED|       72.14 |        2832.66| 0|
(0000)=====2074|CRISTIANO SANTOS ...|PASCULLI|     37| RS |    ACTIVE PERM|ESTABLISHED|       45.87 |        1361.11| 0|
(0000)=====3748|MAISA DA GLORIA V...|FENNIDNO|     52| KJ |    ACTIVE PERM|ESTABLISHED|       37.15 |        2182.37| 0|
(0000)=====3246|CERESA KALIMO MA...|FENNIDNO|     88| RS |    APPENDASOAT |ESTABLISHED|   1822.43 |        5160.92| 0|
(0000)=====3155|SILBERST DO SANT...|PASCULLI|     36| PE |    ACTIVE PERM|ESTABLISHED|   454.200.000.000.000 |         3064.00| 0|
(0000)=====1234|PEDRO DE OLIVEIRA C...|FENNIDNO|     75| GO |    APPENDASOAT |ESTABLISHED|   9722.39 |        33903.32| 0|
(0000)=====7742|JAIZE CARPOS NETTA |FENNIDNO|     55| KJ |    APPENDASOAT |ESTABLISHED| 5.763.229.999.999.999 |         4121.24| 0|
(0000)=====4796|FAEDR DE JESUS RE...|PASCULLI|     52| KJ |    ACTIVE PERM|ESTABLISHED|   559.29 |        321.91| 0|
(0000)=====1234|OTONIE ALVARDO DE ...|PASCULLI|     56| SC |    NO ZAMPHADO |ESTABLISHED|       128 |          306.8 | 0|
```

```
Number of rows in CSV: 6172
24/07/01 18:02:55 INFO SparkBigQueryConnectorModule: Registering cleanup jobs listener, should happen just once
24/07/01 18:03:01 INFO BigQueryDirectDataSoucliriterContext: BigQueryDataSource writer b7b42c0c-096e-4194-a282-32e49d30455 committed with messages:
[BigQueryConnectorClientMessage{partitionId=0, taskId=0, epochId=1719859977694, tableId='projects/halogen-antenna-426018-b0/datasets/CREDITO_CONSIGNADO/tables/tbl_consignado_raw'}]
24/07/01 18:03:02 INFO BigQueryDirectDataSoucliriterContext: BigQueryDataSource writer has committed at time: seconds: 1719859982
nanos: 797007000
```

[illegible]

| id | cpf | nome | seio_saldo | uf | situacao_funcional | tipo_contrato | pmf_valor | margem_saldo | conceder | faixa_10_20 | faixa_21_30 | faixa_31_40 | faixa_41_50 | faixa_51_60 | faixa_61_70 | faixa_71_80 | faixa_81_90 |
|-------------|-------------|-----------------------|------------|----|--------------------|-----------------------|-------------|--------------------|--------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| 00000003287 | 00000003287 | CHARLES ALBERTO S ... | PERNANBU | 76 | AN | APROVEITADO | ESTAVEL | 131.408.999.999-99 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 00000003716 | 00000003716 | ARAJA PAIXA DE ... | PERNANBU | 37 | PE | COLLETTISTA EMPREGADO | ESTAVEL | 7088.7101 | 3184.280 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 00000004134 | 00000004134 | ELINE OPIES DA ... | PERNANBU | 65 | 61 | ATIVO PERMANENTE | ESTAVEL | 2644.421 | 31.728.000.000-00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 00000004368 | 00000004368 | DANIEL ROCHA DO ... | PERNANBU | 38 | PE | ATIVO PERMANENTE | ESTAVEL | 4078.421 | 18233.444 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 00000004712 | 00000004712 | SUZANA DE MATOS ... | PERNANBU | 84 | NO | REEMBOLSADO CONTRATO | NÃO ESTAVEL | 1564.825 | 381.751 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 00000004833 | 00000004833 | EDMUNO GOMES ... | PERNANBU | 81 | PE | ATIVO PERMANENTE | ESTAVEL | 137.091 | 1.084 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 00000004918 | 00000004918 | ANTONIO VIEIRA V ... | PERNANBU | 76 | PE | APROVEITADO | ESTAVEL | 137.091 | 1579.51 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 00000004248 | 00000004248 | CHARLES SARILAND ... | PERNANBU | 31 | AC | ATIVO PERMANENTE | ESTAVEL | 327.15 | 27.941.499.999-99 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 00000004387 | 00000004387 | JOSE ALVES SARA ... | PERNANBU | 52 | PA | ATIVO PERMANENTE | ESTAVEL | 3085.15 | 4135.17 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 00000005058 | 00000005058 | FABRIANO DRECCO ... | PERNANBU | 40 | NO | COLLETTISTA EMPREGADO | ESTAVEL | 8137.81 | 5613.180 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 00000005088 | 00000005088 | HENRIQUE CALISTO ... | PERNANBU | 48 | PE | ATIVO PERMANENTE | ESTAVEL | 8259.8 | 8.999.679.999-99 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 00000005288 | 00000005288 | RENILDA FLORENCE ... | PERNANBU | 81 | NA | ATIVO PERMANENTE | ESTAVEL | 2199.21 | 2199.21 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 00000002975 | 00000002975 | MARIA HELENA CO ... | PERNANBU | 74 | SP | REEMBOLSADO CONTRATO | NÃO ESTAVEL | 1847.919 | 3477 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 00000005881 | 00000005881 | LEONARDO MORAES ... | PERNANBU | 81 | PE | ATIVO PERMANENTE | ESTAVEL | 2081.771 | 2081.771 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 00000008817 | 00000008817 | SARA JOSE SOARES ... | PERNANBU | 37 | TO | ATIVO PERMANENTE | ESTAVEL | 2386.193 | 2097.360 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 00000011755 | 00000011755 | BETA DE CASSIA CO ... | PERNANBU | 53 | NO | REEMBOLSADO CONTRATO | NÃO ESTAVEL | 1577.42 | 15.647.999.999-99 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 00000005253 | 00000005253 | ROSELIANA DO CARO ... | PERNANBU | 78 | PE | ATIVO PERMANENTE | ESTAVEL | 134.361.000.000-00 | 138.781.000.000-00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 00000004817 | 00000004817 | JOSE NEVES REIS ... | PERNANBU | 83 | PE | ATIVO PERMANENTE | ESTAVEL | 13.530.300.000-00 | 3023.94 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 00000056718 | 00000056718 | CATIA REGINA CAU ... | PERNANBU | 54 | NO | REEMBOLSADO CONTRATO | NÃO ESTAVEL | 14.958.100.000-00 | 15.649.799.999-99 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 00000000000 | 00000000000 | PATRICIA HILBERT ... | PERNANBU | 81 | PE | ATIVO PERMANENTE | ESTAVEL | 2086.64 | 2278.0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

```
24/07/01 18:03:11 INFO GoogleCloudStorageFileSystemImpl: Successfully repaired 'gs://halogen-antenna-426018-b0/Prata/consig_transform.parquet/' directory.
```

| | | | | | | | | | | |
|----|----|--------------------|---|---|----|----|-----|-----|-----|----|
| 6 | HT | 262231.06 | 0 | 0 | 55 | 4 | 4 | 4 | 10 | 13 |
| 7 | SP | 1167528.04 | 1 | 2 | 48 | 28 | 18 | 142 | 78 | 30 |
| 8 | P8 | 102356.12000000005 | 0 | 1 | 9 | 59 | 9 | 6 | 14 | 9 |
| 9 | BA | 709931.0599999998 | 0 | 0 | 67 | 58 | 10 | 10 | 38 | 52 |
| 10 | SE | 255462.67000000007 | 0 | 0 | 35 | 19 | 1 | 2 | 6 | 1 |
| 11 | RJ | 1438036.2299999997 | 0 | 1 | 15 | 57 | 681 | 130 | 120 | 64 |
| 12 | AC | 54115.05 | 0 | 8 | 11 | 1 | 2 | 1 | 5 | 3 |
| 13 | PR | 525079.91000000001 | 0 | 1 | 26 | 17 | 3 | 7 | 34 | 42 |
| 14 | AP | 154542.62000000002 | 0 | 6 | 22 | 0 | 1 | 0 | 18 | 6 |
| 15 | RR | 35018.29 | 0 | 1 | 19 | 8 | 1 | 1 | 1 | 0 |
| 16 | TO | 151410.72 | 0 | 1 | 36 | 8 | 4 | 0 | 2 | 1 |
| 17 | ES | 217322.52000000002 | 0 | 0 | 1 | 8 | 27 | 5 | 6 | 5 |
| 18 | AL | 109849.69 | 0 | 0 | 5 | 16 | 4 | 5 | 1 | 8 |
| 19 | RN | 100136.76000000001 | 0 | 0 | 10 | 26 | 3 | 2 | 5 | 3 |

only showing top 20 rows

```
24/07/01 18:03:19 INFO BigQueryDirectDataSourceWriterContext: BigQuery DataSource writer: 944c1cf39-86d3-43da-ef52-82e2607da379 has committed with messages:
[BigQueryWriteCommitInfo{partitionId=0, taskId=0, epochId=1719856989324, tableId='projects/1718601000000/tables/tbl_fato'}}]
24/07/01 18:03:20 INFO BigQueryDirectDataSourceWriterContext: BigQuery DataSource writer has committed at time: seconds: 1719857000
nanos: 672895900
```

```
24/07/01 18:03:20 INFO SparkBigQueryConnectorModule: In SparkListener.onApplicationEnd, going to activate cleanup jobs
```

```
24/07/01 18:03:20 INFO SparkBigQueryConnectorModule: In SparkListener.onAppin
24/07/01 18:03:20 INFO BigQueryClient: Running cleanup jobs. Jobs count is 0
```

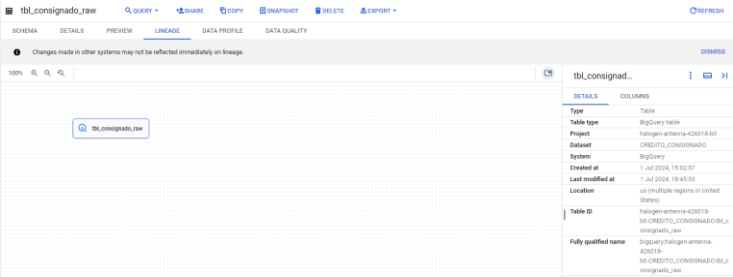
```
24/07/01 18:03:20 INFO BigQueryClient: Running cleanup jobs. Jobs count is 0
24/07/01 18:03:20 INFO BigQueryClient: Clearing the cleanup jobs list
```

```
24/07/01 18:03:20 INFO BigQueryClient: Finished to run cleanup jobs.
```

| | Job ID | Status | Region | Type | Cluster | Start time | Elapsed time | Labels |
|--|-------------------------------|---|-------------|---------|------------------------------|----------------------|--------------|--------|
| | Pipeline_FULL | Succeeded | us-central1 | PySpark | cluster-6717 | 1 Jul 2024, 15:02:21 | 1 min 1 sec | None |

1.4.3.2. Linhagem

Como os dados não sofreram alteração, por ser uma base real já tratada, não conseguirei detalhamento sobre o tema. Deixarei o print abaixo, como forma de representação do controle no bigquery:



1.4.3.3. Dataplex (Data Quality)

Utilizamos o Dataplex para realização do Data Quality. Foram adicionadas regras de negócio para validação dos campos do dataset original. Importante lembrar que os dados de origem já estavam previamente tratados, por se tratar de uma base privada.

Configuração do Dataplex com as regras de negócio:

| Filter Filter items | | | | | | | | |
|--------------------------|--------------|-----------|---------------------|------------|--------------|--|-----------|--|
| <input type="checkbox"/> | Column name | Rule name | Rule type | Evaluation | Dimension | Parameters | Threshold | Actions |
| <input type="checkbox"/> | conceder | - | Value Set Check | Per row | Validity | set of: 1,0 | 100% | <input type="checkbox"/> <input checked="" type="checkbox"/> |
| <input type="checkbox"/> | cpf | - | Null Check | Per row | Completeness | | 100% | <input type="checkbox"/> <input checked="" type="checkbox"/> |
| <input type="checkbox"/> | cpf | - | Row Condition Check | Per row | Validity | (LENGTH('cpf') >= 11 AND LENGTH('cpf') <= 11) OR 'cpf' IS NULL | 100% | <input type="checkbox"/> <input checked="" type="checkbox"/> |
| <input type="checkbox"/> | idade | - | Null Check | Per row | Completeness | | 100% | <input type="checkbox"/> <input checked="" type="checkbox"/> |
| <input type="checkbox"/> | margem_saldo | - | Null Check | Per row | Completeness | | 100% | <input type="checkbox"/> <input checked="" type="checkbox"/> |
| <input type="checkbox"/> | margem_saldo | - | Row Condition Check | Per row | Validity | (LENGTH('margem_saldo') >= 1 AND LENGTH('margem_saldo') <= 23) OR 'margem_saldo' IS NULL | 100% | <input type="checkbox"/> <input checked="" type="checkbox"/> |
| <input type="checkbox"/> | nome | - | Null Check | Per row | Completeness | | 100% | <input type="checkbox"/> <input checked="" type="checkbox"/> |
| <input type="checkbox"/> | pmt_valor | - | Null Check | Per row | Completeness | | 100% | <input type="checkbox"/> <input checked="" type="checkbox"/> |
| <input type="checkbox"/> | pmt_valor | - | Row Condition Check | Per row | Validity | (LENGTH('pmt_valor') >= 1 AND LENGTH('pmt_valor') <= 22) OR 'pmt_valor' IS NULL | 100% | <input type="checkbox"/> <input checked="" type="checkbox"/> |
| <input type="checkbox"/> | sexo | - | Null Check | Per row | Completeness | | 100% | <input type="checkbox"/> <input checked="" type="checkbox"/> |
| <input type="checkbox"/> | sexo | - | Value Set Check | Per row | Validity | set of: FEMININO,MASCULINO | 100% | <input type="checkbox"/> <input checked="" type="checkbox"/> |
| <input type="checkbox"/> | sexo | - | Row Condition Check | Per row | Validity | (LENGTH('sexo') >= 8 AND LENGTH('sexo') <= 9) OR 'sexo' IS NULL | 100% | <input type="checkbox"/> <input checked="" type="checkbox"/> |
| <input type="checkbox"/> | uf | - | Null Check | Per row | Completeness | | 100% | <input type="checkbox"/> <input checked="" type="checkbox"/> |
| <input type="checkbox"/> | uf | - | Row Condition Check | Per row | Validity | (LENGTH('uf') >= 2 AND LENGTH('uf') <= 2) OR 'uf' IS NULL | 100% | <input type="checkbox"/> <input checked="" type="checkbox"/> |

Sucesso na validação do Job:

| Job ID | Start time | Records scanned | Job status | Data quality result |
|--------------------------------------|-------------------------------|-----------------|---|--|
| 02372147-1c30-42a1-b086-1479614841fe | 1 July 2024 at 18:46:24 UTC-3 | 6172 | <input checked="" type="checkbox"/> Succeeded | <input checked="" type="checkbox"/> Passed |

Detalhamento dos testes realizados:

Job ID: 02372147-1c30-42a1-b086-1479614841fe results

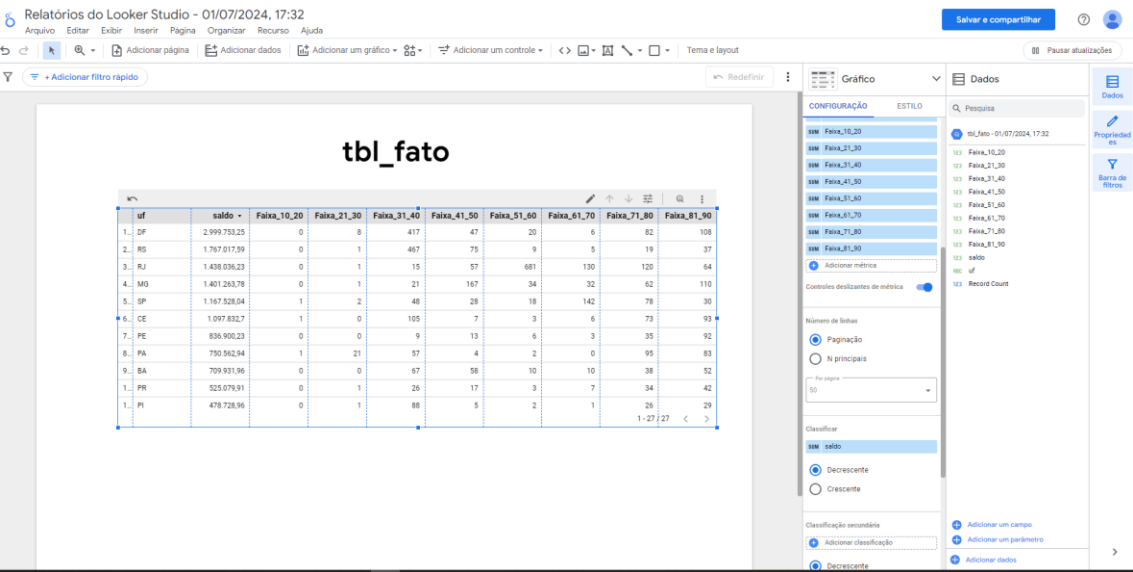
| | | | |
|-----------------------|----------------------|-----------------------|---------------------------|
| Overall Score 100% | 14 PASSED RULES ✔ | Validity 100% ✔ | Completeness 100% ✔ |
|-----------------------|----------------------|-----------------------|---------------------------|

Rules

| Column name | Rule name | Rule type | Status | Evaluation | Dimension | Parameters | Failed rows | Threshold | Query to get failed records |
|--------------|-----------|---------------------|----------|------------|--------------|---------------------------|-------------|-----------|-----------------------------------|
| conceder | - | Value Set Check | ✔ Passed | Per row | Validity | set of: 1,0 | 0% | 100% | WITH '02372147-1c30-42a1-b086-... |
| cpf | - | Null Check | ✔ Passed | Per row | Completeness | | 0% | 100% | WITH '02372147-1c30-42a1-b086-... |
| cpf | - | Row Condition Check | ✔ Passed | Per row | Validity | (LENGTH('cpf') >= 11 AND | 0% | 100% | WITH '02372147-1c30-42a1-b086-... |
| idade | - | Null Check | ✔ Passed | Per row | Completeness | | 0% | 100% | WITH '02372147-1c30-42a1-b086-... |
| margem_saldo | - | Null Check | ✔ Passed | Per row | Completeness | | 0% | 100% | WITH '02372147-1c30-42a1-b086-... |
| margem_saldo | - | Row Condition Check | ✔ Passed | Per row | Validity | (LENGTH('margem_saldo') | 0% | 100% | WITH '02372147-1c30-42a1-b086-... |
| nome | - | Null Check | ✔ Passed | Per row | Completeness | | 0% | 100% | WITH '02372147-1c30-42a1-b086-... |
| pmt_valor | - | Null Check | ✔ Passed | Per row | Completeness | | 0% | 100% | WITH '02372147-1c30-42a1-b086-... |
| pmt_valor | - | Row Condition Check | ✔ Passed | Per row | Validity | (LENGTH('pmt_valor') >= 1 | 0% | 100% | WITH '02372147-1c30-42a1-b086-... |
| sexo | - | Null Check | ✔ Passed | Per row | Completeness | | 0% | 100% | WITH '02372147-1c30-42a1-b086-... |
| sexo | - | Value Set Check | ✔ Passed | Per row | Validity | set of: | 0% | 100% | WITH '02372147-1c30-42a1-b086-... |

1.4.4. Looker Stúdio (Análises)

O Looker Stúdio é uma ferramenta de análise acoplada ao Bigquery.



1.4.4.1. Volume de idade por UF

Neste item, precisamos entender se faz sentido segmentarmos a base pela UF e grupo de faixas de atrasos.

Baseado no gráfico abaixo, percebemos nitidamente uma aglomeração muito alta na faixa de 31_40, para minha "surpresa". Seguidos pela faixa de 81_90, 51_60 e 71_80.

Consulta:

```
SELECT
  distinct(uf) as uf,
  Margem_Saldo as saldo,
  Faixa_10_20,
  Faixa_21_30,
  Faixa_31_40,
  Faixa_41_50,
  Faixa_51_60,
  Faixa_61_70,
  Faixa_71_80,
  Faixa_81_90
FROM `halogen-antenna-426018-b0.CREDITO_CONSIGNADO.tbl_fato`
--group by uf
order by Margem_Saldo desc
```

Execução no Bigquery:

Query results

| Row | uf | Faixa_10_20 | Faixa_21_30 | Faixa_31_40 | Faixa_41_50 | Faixa_51_60 | Faixa_61_70 | Faixa_71_80 | Faixa_81_90 |
|-----|----|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| 1 | MS | 1 | 1 | 84 | 5 | 4 | 5 | 3 | 4 |
| 2 | CE | 1 | 0 | 105 | 7 | 3 | 6 | 73 | 93 |
| 3 | MG | 0 | 1 | 21 | 167 | 34 | 22 | 62 | 110 |
| 4 | DF | 0 | 8 | 417 | 47 | 20 | 6 | 82 | 108 |
| 5 | RO | 1 | 13 | 21 | 1 | 1 | 2 | 4 | 0 |
| 6 | AM | 1 | 13 | 37 | 6 | 0 | 0 | 42 | 34 |
| 7 | MT | 0 | 0 | 55 | 4 | 4 | 4 | 10 | 13 |
| 8 | SP | 1 | 2 | 48 | 28 | 18 | 142 | 78 | 30 |
| 9 | PB | 0 | 1 | 9 | 59 | 9 | 6 | 14 | 9 |
| 10 | BA | 0 | 0 | 67 | 58 | 10 | 10 | 38 | 52 |
| 11 | SE | 0 | 0 | 35 | 19 | 1 | 2 | 6 | 1 |
| 12 | RJ | 0 | 1 | 15 | 57 | 681 | 130 | 120 | 64 |
| 13 | AC | 0 | 8 | 11 | 1 | 2 | 1 | 5 | 3 |
| 14 | PR | 0 | 1 | 26 | 17 | 3 | 7 | 34 | 42 |
| 15 | AP | 0 | 6 | 22 | 0 | 1 | 0 | 18 | 6 |

Results per page: 50 1 - 27 of 27

| | uf | Faixa_10_20 | Faixa_21_30 | Faixa_31_40... | Faixa_51_60 | Faixa_61_70 | Faixa_71_80 | Faixa_81_90 |
|-----|-------------|-------------|-------------|----------------|-------------|-------------|-------------|-------------|
| 1. | RS | 0 | 1 | 467 | 9 | 5 | 19 | 37 |
| 2. | DF | 0 | 8 | 417 | 20 | 6 | 82 | 108 |
| 3. | GO | 0 | 3 | 125 | 4 | 2 | 6 | 5 |
| 4. | CE | 1 | 0 | 105 | 3 | 6 | 73 | 93 |
| 5. | PI | 0 | 1 | 88 | 2 | 1 | 26 | 29 |
| 6. | MS | 1 | 1 | 84 | 4 | 5 | 3 | 4 |
| 7. | MA | 1 | 0 | 76 | 2 | 2 | 16 | 28 |
| 8. | BA | 0 | 0 | 67 | 10 | 10 | 38 | 52 |
| 9. | PA | 1 | 21 | 57 | 2 | 0 | 95 | 83 |
| 10. | MT | 0 | 0 | 55 | 4 | 4 | 10 | 13 |
| 11. | SP | 1 | 2 | 48 | 18 | 142 | 78 | 30 |
| 12. | AM | 1 | 13 | 37 | 0 | 0 | 42 | 34 |
| 13. | TO | 0 | 1 | 36 | 4 | 0 | 2 | 1 |
| 14. | SE | 0 | 0 | 35 | 1 | 2 | 6 | 1 |
| 15. | SC | 0 | 0 | 32 | 10 | 5 | 12 | 14 |
| 16. | PR | 0 | 1 | 26 | 3 | 7 | 34 | 42 |
| 17. | AP | 0 | 6 | 22 | 1 | 0 | 18 | 6 |
| 18. | RO | 1 | 13 | 21 | 1 | 2 | 4 | 0 |
| 19. | MG | 0 | 1 | 21 | 34 | 32 | 62 | 110 |
| 20. | RR | 1 | 8 | 19 | 1 | 1 | 1 | 0 |
| | Total geral | 8 | 90 | 1.898 | 865 | 384 | 813 | 874 |

Já em relação a UF, notamos como aglomeração nas UFs: RS,DF,GO,CE para faixa 31_40:

| | uf | Faixa_10_20 | Faixa_21_30 | Faixa_31_40... | Faixa_51_60 | Faixa_61_70 | Faixa_71_80 | Faixa_81_90 |
|-----|-------------|-------------|-------------|----------------|-------------|-------------|-------------|-------------|
| 1. | RS | 0 | 1 | 467 | 9 | 5 | 19 | 37 |
| 2. | DF | 0 | 8 | 417 | 20 | 6 | 82 | 108 |
| 3. | GO | 0 | 3 | 125 | 4 | 2 | 6 | 5 |
| 4. | CE | 1 | 0 | 105 | 3 | 6 | 73 | 93 |
| 5. | PI | 0 | 1 | 88 | 2 | 1 | 26 | 29 |
| 6. | MS | 1 | 1 | 84 | 4 | 5 | 3 | 4 |
| 7. | MA | 1 | 0 | 76 | 2 | 2 | 16 | 28 |
| 8. | BA | 0 | 0 | 67 | 10 | 10 | 38 | 52 |
| 9. | PA | 1 | 21 | 57 | 2 | 0 | 95 | 83 |
| 10. | MT | 0 | 0 | 55 | 4 | 4 | 10 | 13 |
| 11. | SP | 1 | 2 | 48 | 18 | 142 | 78 | 30 |
| 12. | AM | 1 | 13 | 37 | 0 | 0 | 42 | 34 |
| 13. | TO | 0 | 1 | 36 | 4 | 0 | 2 | 1 |
| 14. | SE | 0 | 0 | 35 | 1 | 2 | 6 | 1 |
| 15. | SC | 0 | 0 | 32 | 10 | 5 | 12 | 14 |
| 16. | PR | 0 | 1 | 26 | 3 | 7 | 34 | 42 |
| 17. | AP | 0 | 6 | 22 | 1 | 0 | 18 | 6 |
| 18. | RO | 1 | 13 | 21 | 1 | 2 | 4 | 0 |
| 19. | MG | 0 | 1 | 21 | 34 | 32 | 62 | 110 |
| 20. | RR | 1 | 8 | 19 | 1 | 1 | 1 | 0 |
| | Total geral | 8 | 90 | 1.898 | 865 | 384 | 813 | 874 |

Na faixa de 81_90, notamos interseção apenas na UF de DF:

| | uf | Faixa_10_20 | Faixa_21_30 | Faixa_31_40 | Faixa_51_60 | Faixa_61_70 | Faixa_71_80 | Faixa_81_90... |
|-------------|----|-------------|-------------|-------------|-------------|-------------|-------------|----------------|
| 1. | MG | 0 | 1 | 21 | 34 | 32 | 62 | 110 |
| 2. | DF | 0 | 8 | 417 | 20 | 6 | 82 | 108 |
| 3. | CE | 1 | 0 | 105 | 3 | 6 | 73 | 93 |
| 4. | PE | 0 | 0 | 9 | 6 | 3 | 35 | 92 |
| 5. | PA | 1 | 21 | 57 | 2 | 0 | 95 | 83 |
| 6. | RJ | 0 | 1 | 15 | 681 | 130 | 120 | 64 |
| 7. | BA | 0 | 0 | 67 | 10 | 10 | 38 | 52 |
| 8. | PR | 0 | 1 | 26 | 3 | 7 | 34 | 42 |
| 9. | RS | 0 | 1 | 467 | 9 | 5 | 19 | 37 |
| 10. | AM | 1 | 13 | 37 | 0 | 0 | 42 | 34 |
| 11. | SP | 1 | 2 | 48 | 18 | 142 | 78 | 30 |
| 12. | PI | 0 | 1 | 88 | 2 | 1 | 26 | 29 |
| 13. | MA | 1 | 0 | 76 | 2 | 2 | 16 | 28 |
| 14. | SC | 0 | 0 | 32 | 10 | 5 | 12 | 14 |
| 15. | MT | 0 | 0 | 55 | 4 | 4 | 10 | 13 |
| 16. | PB | 0 | 1 | 9 | 9 | 6 | 14 | 9 |
| 17. | AL | 0 | 0 | 5 | 4 | 5 | 1 | 8 |
| 18. | AP | 0 | 6 | 22 | 1 | 0 | 18 | 6 |
| 19. | GO | 0 | 3 | 125 | 4 | 2 | 6 | 5 |
| 20. | ES | 0 | 0 | 1 | 27 | 5 | 6 | 5 |
| Total geral | | 8 | 90 | 1.898 | 865 | 384 | 813 | 874 |

1 - 27 / 27 < >

1.4.4.2. Volume da margem de saldo por UF

Neste item, precisamos entender se faz sentido segmentarmos a base pela UF e margem de saldo. Através do gráfico abaixo, percebemos que DF, RS, RJ, MG, SP e CE possuem um valor financeiro considerável.

Consulta:

```
SELECT
  distinct(uf) as uf,
  Margem_Saldo as saldo,
FROM `halogen-antenna-426018-b0.CREDITO_CONSIGNADO.tbl_fato`
order by Margem_Saldo desc
```

Execução no Bigquery:

Untitled query

RUN

SAVE

DOWNLOAD

SHARE

SCHEDULE

MORE

1

SELECT

2

distinct(uf) as uf,

3

Margem_Saldo as saldo,

4

FROM `halogen-antenna-426018-b0.CREDITO_CONSIGNADO.tbl_fato`

5

--group by uf

6

order by Margem_Saldo desc

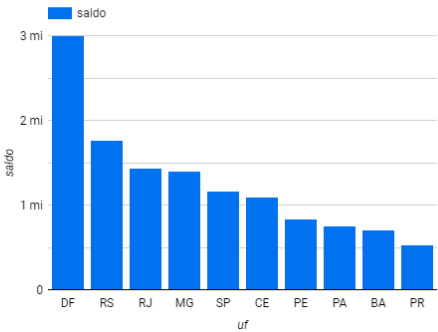
Query results

| JOB INFORMATION | | RESULTS | CHART | JSON | EXECUTION DETAILS | EXECUTION GRAPH |
|-----------------|----|-------------------|-------|------|-------------------|-----------------|
| Row | uf | saldo | | | | |
| 1 | DF | 2999753.249999... | | | | |
| 2 | RS | 1767017.589999... | | | | |
| 3 | RJ | 1438036.229999... | | | | |
| 4 | MG | 1401263.780000... | | | | |
| 5 | SP | 1167528.04 | | | | |
| 6 | CE | 1097832.699999... | | | | |
| 7 | PE | 836900.229999... | | | | |
| 8 | PA | 750562.939999... | | | | |
| 9 | BA | 709931.959999... | | | | |
| 10 | PR | 525079.910000... | | | | |
| 11 | PI | 478728.959999... | | | | |
| 12 | SC | 462704.87 | | | | |
| 13 | GO | 438537.560000... | | | | |
| 14 | AM | 430345.589999... | | | | |
| 15 | MA | 408980.860000... | | | | |

tbl_fato

| | uf | saldo |
|-----|----|---------------|
| 1. | DF | 2.999.753,... |
| 2. | RS | 1.767.017,... |
| 3. | RJ | 1.438.036,... |
| 4. | MG | 1.401.263,... |
| 5. | SP | 1.167.528,... |
| 6. | CE | 1.097.832,7 |
| 7. | PE | 836.900,23 |
| 8. | PA | 750.562,94 |
| 9. | BA | 709.931,96 |
| 10. | PR | 525.079,91 |

1 - 27 / 27 < >



1.4.4.3. Volume da margem de saldo por UF e grupo de idade

Neste último item, precisamos entender se faz sentido segmentarmos a base pelo saldo, uf e grupo de idade. Utilizamos como referência a faixa de 31_40, devido ao alto volume encontrado no item 1.4.4.1, deste documento.

Consulta:

```
SELECT
  distinct(uf) as uf,
  Margem_Saldo as saldo,
  Faixa_10_20,
  Faixa_21_30,
  Faixa_31_40,
  Faixa_41_50,
  Faixa_51_60,
  Faixa_61_70,
  Faixa_71_80,
  Faixa_81_90
FROM `halogen-antenna-426018-b0.CREDITO_CONSIGNADO.tbl_fato`
order by Margem_Saldo desc
```

Execução no Bigquery:

Query results

1 SELECT
2 distinct(uf) as uf,
3 Margem_Saldo,
4 Faixa_10_20,
5 Faixa_21_30,
6 Faixa_31_40,
7 Faixa_41_50,
8 Faixa_51_60,
9 Faixa_61_70,
10 Faixa_71_80,
11 Faixa_81_90
12 FROM `halogen-antenna-426018-b0.CREDITO_CONSIGNADO.tbl_fato`
13 order by Margem_Saldo desc

Query results

SAVE RESULTS

| Row | uf | Margem_Saldo | Faixa_10_20 | Faixa_21_30 | Faixa_31_40 | Faixa_41_50 | Faixa_51_60 | Faixa_61_70 | Faixa_71_80 | Faixa_81_90 |
|-----|----|-------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| 1 | DF | 2999753.249999... | 0 | 8 | 417 | 47 | 20 | 6 | 82 | 108 |
| 2 | RS | 1767017.589999... | 0 | 1 | 467 | 75 | 9 | 5 | 19 | 37 |
| 3 | RJ | 1438036.229999... | 0 | 1 | 15 | 57 | 681 | 130 | 120 | 64 |
| 4 | MG | 1401263.780000... | 0 | 1 | 21 | 167 | 34 | 32 | 62 | 110 |
| 5 | SP | 1167528.04 | 1 | 2 | 48 | 28 | 18 | 142 | 78 | 30 |
| 6 | CE | 1097832.699999... | 1 | 0 | 105 | 7 | 3 | 6 | 73 | 93 |
| 7 | PE | 836900.229999... | 0 | 0 | 9 | 13 | 6 | 3 | 35 | 92 |
| 8 | PA | 750562.929999... | 1 | 21 | 57 | 4 | 2 | 0 | 95 | 83 |
| 9 | BA | 709931.959999... | 0 | 0 | 67 | 58 | 10 | 10 | 38 | 52 |
| 10 | PR | 525079.910000... | 0 | 1 | 26 | 17 | 3 | 7 | 34 | 42 |
| 11 | PI | 478728.959999... | 0 | 1 | 88 | 5 | 2 | 1 | 26 | 29 |
| 12 | SC | 462704.87 | 0 | 0 | 32 | 26 | 10 | 5 | 12 | 14 |
| 13 | GO | 438537.560000... | 0 | 3 | 125 | 9 | 4 | 2 | 6 | 5 |
| 14 | AM | 430345.589999... | 1 | 13 | 37 | 6 | 0 | 0 | 42 | 34 |

Através do gráfico abaixo, percebemos que RS, DF, GO e CE como UFs relevantes.

| uf | saldo | Faixa_10_20 | Faixa_21_30 | Faixa_31_4... | Faixa_41_50 | Faixa_51_60 | Faixa_61_70 | Faixa_71_80 | Faixa_81_90 |
|--------|--------------|-------------|-------------|---------------|-------------|-------------|-------------|-------------|-------------|
| 1.. RS | 1.767.017,59 | 0 | 1 | 467 | 75 | 9 | 5 | 19 | 37 |
| 2.. DF | 2.999.753,25 | 0 | 8 | 417 | 47 | 20 | 6 | 82 | 108 |
| 3.. GO | 438.537,56 | 0 | 3 | 125 | 9 | 4 | 2 | 6 | 5 |
| 4.. CE | 1.097.832,7 | 1 | 0 | 105 | 7 | 3 | 6 | 73 | 93 |
| 5.. PI | 478.728,96 | 0 | 1 | 88 | 5 | 2 | 1 | 26 | 29 |
| 6.. MS | 200.709,97 | 1 | 1 | 84 | 5 | 4 | 5 | 3 | 4 |
| 7.. MA | 408.980,86 | 1 | 0 | 76 | 5 | 2 | 2 | 16 | 28 |
| 8.. BA | 709.931,96 | 0 | 0 | 67 | 58 | 10 | 10 | 38 | 52 |
| 9.. PA | 750.562,94 | 1 | 21 | 57 | 4 | 2 | 0 | 95 | 83 |
| 1.. MT | 262.231,06 | 0 | 0 | 55 | 4 | 4 | 4 | 10 | 13 |
| 1.. SP | 1.167.528,04 | 1 | 2 | 48 | 28 | 18 | 142 | 78 | 30 |

1.4.5. Conclusão

1.4.5.1. Considerações

A faixa de 31 à 40 anos se destaca em termos de volume e margem de saldo, indicando um potencial interessante para oferta do crédito consignado. Confesso que fiquei muito surpreso, pois não tinha essa visão. Concentrava grande parte das ações nas faixas superiores à 50 anos.

As UF: RS,DS,GO,CE apresentam concentração significativa na faixa etária desejada e com margem de saldo favorável, tornando as áreas prioritárias para prospecção.

A análise detalhada por UF e faixa etária permite identificar oportunidades específicas para cada região, otimizando as estratégias de venda.

1.4.5.2. Recomendações

- Focar na faixa etária de 31 a 40 anos: Direcionar esforços de marketing e vendas para este público, que apresenta maior potencial de conversão.
- Priorizar as UFs RS, DF, GO e CE: Concentrar ações nestas regiões, onde existe maior concentração de clientes com perfil adequado ao crédito consignado.
- Analisar dados por UF e faixa etária: Segmentar as análises para identificar oportunidades específicas em cada região e faixa etária, personalizando as ofertas e otimizando os resultados.
- Monitorar indicadores de performance: Acompanhar métricas como taxa de aprovação, volume de concessões e inadimplência para avaliar a efetividade das estratégias e realizar ajustes quando necessário.