

Curso: Ciência de Dados e Analytics

Sprint: Engenharia de Dados

Aluno: Henrique Meirelles Francisco

Objetivo

O **Desenrola Brasil** é o Programa de Renegociação de Créditos Inadimplidos, criado pelo Governo Federal, com o objetivo de recuperar as condições de crédito de Devedores que possuam dívidas negativadas.

As negociações são feitas totalmente por ***meio digital***, com uma navegação intuitiva e rápida, garantindo agilidade, comodidade e conveniência para a regularização dos seus débitos.

Os dados obtidos do programa do governo federal: Desenrola Brasil, contém dados relacionados às operações financeiras de vários conglomerados financeiros. Inclui detalhes como o número de operações, seu volume, conglomerado e Unidades da federação.

Estes dados serão usados para os seguintes indicadores neste primeiro MVP:

1. Percentual de operações por Unidade da Federação;
2. Volume e percentual de Operações por Conglomerado Financeiro (algum banco 100% digital se destacará)?
3. Volume de operações por data base;

Detalhamento

Os dados utilizados no MVP foram obtidos do portal de dados abertos do Banco Central do Brasil e carregados no DataBricks. A modelagem dos dados resultou em uma única entidade chamada "Dados Desenrola", que inclui informações como número de operações, volume de operações, conglomerado financeiro e unidade da federação.

1. Busca pelos dados

- Foi utilizado o portal de dados abertos oferecido pelo Banco Central do Brasil.

<https://dadosabertos.bcb.gov.br/dataset>

The screenshot displays the 'dadosabertos.bcb.gov.br/dataset' page. The header includes the 'gov.br' logo and navigation links: 'ACESSO À INFORMAÇÃO', 'PARTICIPE', 'LEGISLAÇÃO', and 'ÓRGÃOS DO GOVERNO'. The main navigation bar features 'Dados', 'Perguntas Frequentes', 'Contato', and 'Sobre o Portal'. A breadcrumb trail shows 'Conjuntos de dados'. On the left, a sidebar lists various organizations and groups with their respective dataset counts. The main content area shows a search bar with the text 'Pesquisar conjuntos de dados...', a search button, and a result count of '4.046 conjuntos de dados encontrados'. Below this, several dataset entries are listed, each with a title, description, and download links for HTML, API, and JSON formats. The datasets include 'SFN - WESTERN ASSET MANAGEMENT COMPA', 'SFN - RENDIMENTO DTVM S.A.', 'SFN - BCO SANTANDER (BRASIL) S.A.', and 'SFN - UNICRED DISTRIBUIDORA DE TÍTULO'. At the bottom, a link for 'Taxa de juros - Selic anualizada base 252' is visible.

Dentro deste site localizamos os dados do BCB/Desig – Departamento de Monitoramento do Sistema Financeiro.

- A base de dados utilizada para este MVP vem do programa de governo Desenrola Brasil.

https://www.bcb.gov.br/pda/desig/desenrola/dados_desenrola.csv

https://dadosabertos.bcb.gov.br/dataset/desenrola-brasil/resource/ec74fbf9-fcce-4166-91c1-228eaa3cde5f?view_id=2281721d-c802-4d19-b792-a575539ff626

gouvbr

ACESSO À INFORMAÇÃO PARTICIPE LEGISLAÇÃO ÓRGÃOS DO GOVERNO

BANCO CENTRAL DO BRASIL

Dados Perguntas Frequentes Contato Sobre o Portal

> Organizações > BCB/Desig > Desenrola Brasil > Desenrola Brasil

Desenrola Brasil

URL: https://www.bcb.gov.br/pda/desig/desenrola/dados_desenrola.csv

Baixar todos os dados no formato CSV.

Explorador de Dados

Tela cheia Embutir

Grade Gráfico Mapa 1000 records 1 - 100



Search data ... Go Filtros

DATA_B...
202309;2...
202309;2...
202309;2...
202309;2...
202309;2...
202309;2...

2. Coleta

O arquivo ***dados_desenrola.csv*** foi copiado para máquina local e depois carregado para dentro do workspace no DataBricks.

Diretório local:

Nome	Status	Data	Tipo	Tamanho	Marcas
 dados_desenrola.csv		31/03/2025 23:33	Arquivo de Valore...	324 KB	

Workspace no databricks:

← → ↻ 🔒 https://dbc-b7a92467-5d84.cloud.databricks.com/workspace/760444652908315?o=1118165595641490

Search data, notebooks, recents, and more... CTRL + P

Unlock account

+ New

Workspace

Recents

Catalog

Workflows

Compute

Marketplace

SQL

SQL Editor

Queries

Dashboards

Genie

Alerts

Query History

SQL Warehouses

Data Engineering

Job Runs

Data Ingestion

Pipelines

Machine Learning

Playground

Experiments

Features

dados_desenrola.csv

File Edit View Run Help Last edit was 5 minutes ago

Workspace

henrique.mfrancis@gmail.com

create-pipeline-from-sample-data

Henrique_dados_desenrola2

Bakehouse Sales Starter Space

dados_desenrola 2025-03-31 22:19:35

dados_desenrola.csv

MVP 1 - Henrique Meirelles Francisco

MVP1 - Query

Untitled Notebook 2025-03-31 21:34:33

Workspace Usage Dashboard

```
1 DATA_BASE;TIPO_DESENROLA;UNIDADE_FEDERACAO;COD_CONGLOMERADO_FINANCEIRO;NOME_CONGLOMERADO_FINANCEIRO;NUMERO_OPERACOES;  
2 VOLUME_OPERACOES  
3 202309;2;AC;49906;BB;142;1421991,1  
4 202309;2;AC;10045;BRADESCO;17;94097,52  
5 202309;2;AC;49944;BTG PACTUAL;7;37788,79  
6 202309;2;AC;51626;CAIXA ECONOMICA FEDERAL;82;247430,66  
7 202309;2;AC;51884;INTER;4;12306,86  
8 202309;2;AC;10069;ITAU;148;327784,69  
9 202309;2;AC;30379;SANTANDER;232;1025040,23  
10 202309;2;AC;51811;VOTORANTIM;17;64294,81  
11 202309;2;AC;49906;BB;366;1915287,58  
12 202309;2;AL;13009717;BCO DO EST. DE SE S.A.;1;3362,67  
13 202309;2;AL;10045;BRADESCO;106;504918,94  
14 202309;2;AL;49944;BTG PACTUAL;21;132834,28  
15 202309;2;AL;51626;CAIXA ECONOMICA FEDERAL;313;693114,53  
16 202309;2;AL;51884;INTER;34;131191,36  
17 202309;2;AL;10069;ITAU;1374;2988789,99  
18 202309;2;AL;30379;SANTANDER;1269;4077628,9  
19 202309;2;AL;51811;VOTORANTIM;47;148416,15  
20 202309;2;AM;49906;BB;314;2570440,44  
21 202309;2;AM;10045;BRADESCO;369;2476907,16  
22 202309;2;AM;49944;BTG PACTUAL;21;153686  
23 202309;2;AM;51626;CAIXA ECONOMICA FEDERAL;219;437871,38  
24 202309;2;AM;51884;INTER;23;75962,24  
25 202309;2;AM;10069;ITAU;1314;3176353,7  
26 202309;2;AM;30379;SANTANDER;2044;8604597,83  
27 202309;2;AP;49906;BB;225;2115839,66  
28 202309;2;AP;10045;BRADESCO;19;188133,81  
29 202309;2;AP;49944;BTG PACTUAL;6;34749,99  
30 202309;2;AP;51626;CAIXA ECONOMICA FEDERAL;64;148091,84  
31 202309;2;AP;51884;INTER;14;48998,53  
32 202309;2;AP;10069;ITAU;203;482458,2
```

3. Modelagem

O modelo contém uma única entidade:

Dados Desenrola

Dados Desenrola
<ul style="list-style-type: none">• DATA_BASE• TIPO_DESENROLA• UNIDADE_FEDERAÇÃO• COD_CONGLOMERADO_FINANCEIRO• NOME_CONGLOMERADO_FINANCEIRO• NÚMERO_OPERAÇÕES• VOLUME_OPERAÇÕES

Dicionário de dados:

Nome da Coluna	Tipo	Descrição
DATA_BASE	STRING	Mês de referência no formato AAAAMM.
TIPO_DESENROLA	INT	Tipos 1 e 2, correspondendo às faixas 1 e 2 ⁽¹⁾ , e tipo 3 ⁽²⁾ .
UNIDADE_FEDERACAO	STRING	Sigla da unidade da federação.
COD_CONGLOMERADO_FINANCEIRO	INT	Código do conglomerado financeiro.
NOME_CONGLOMERADO_FINANCEIRO	STRING	Nome do conglomerado financeiro.
NUMERO_OPERACOES	INT	Número de operações renegociadas no mês de referência.
VOLUME_OPERACOES	DOUBLE	Somatório dos valores das operações após a concessão do desconto, em reais, renegociadas no mês de referência (casa decimal separada por vírgula).

⁽¹⁾, ⁽²⁾: As Faixas 1, 2 e 3 do Desenrola Brasil são etapas do programa de renegociação de dívidas, que foi criado pelo Governo Federal. O objetivo é ajudar as pessoas a recuperar o crédito e voltar a ter condições de adquirir novos empréstimos.

Faixa 1:

- Destinada a quem recebe até dois salários-mínimos ou está inscrito no CadÚnico;
- Dívidas bancárias ou não bancárias (conta de água e luz, por exemplo) de até R\$ 20 mil;
- Negativadas entre janeiro de 2019 e dezembro de 2022.

Faixa 2:

- Renegociações realizadas diretamente com os bancos credores;
- Destinada a pessoas físicas com renda bruta mensal de até R\$ 20 mil.

Faixa 3

- Renegociação de dívidas com empresas de diversos setores, como comércio varejista, eletricidade, telecomunicações, educação, saneamento, serviços financeiros, securitizadoras, micro e pequenas empresas.

4. Carga

a. Criação da tabela `dados_desenrola` no catálogo `workspace`.

Notebook MVP 1 - Henrique Meirelles Francisco:

<https://dbc-b7a924d7-5c84.cloud.databricks.com/editor/notebooks/3535255726529390?o=1118165595641490>

The screenshot shows a Databricks SQL notebook interface. The left sidebar contains navigation options like Workspace, Recents, Catalog, Workflows, Compute, Marketplace, SQL, SQL Editor, Queries, Dashboards, Genie, Alerts, Query History, and SQL Warehouses. The main area displays a SQL query: `USE CATALOG 'workspace'; CREATE TABLE IF NOT EXISTS default.dados_desenrola (DATA_BASE STRING, TIPO_DESENROLA INT, UNIDADE_FEDERACAO STRING, COD_CONGLOMERADO_FINANCEIRO INT, NOME_CONGLOMERADO_FINANCEIRO STRING, NUMERO_OPERACOES INT, VOLUME_OPERACOES DOUBLE);`. Below the query, a performance table is visible:

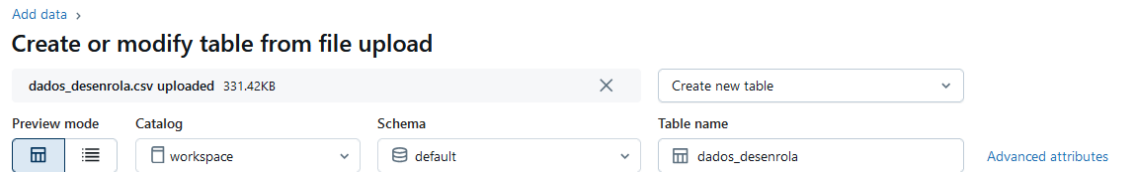
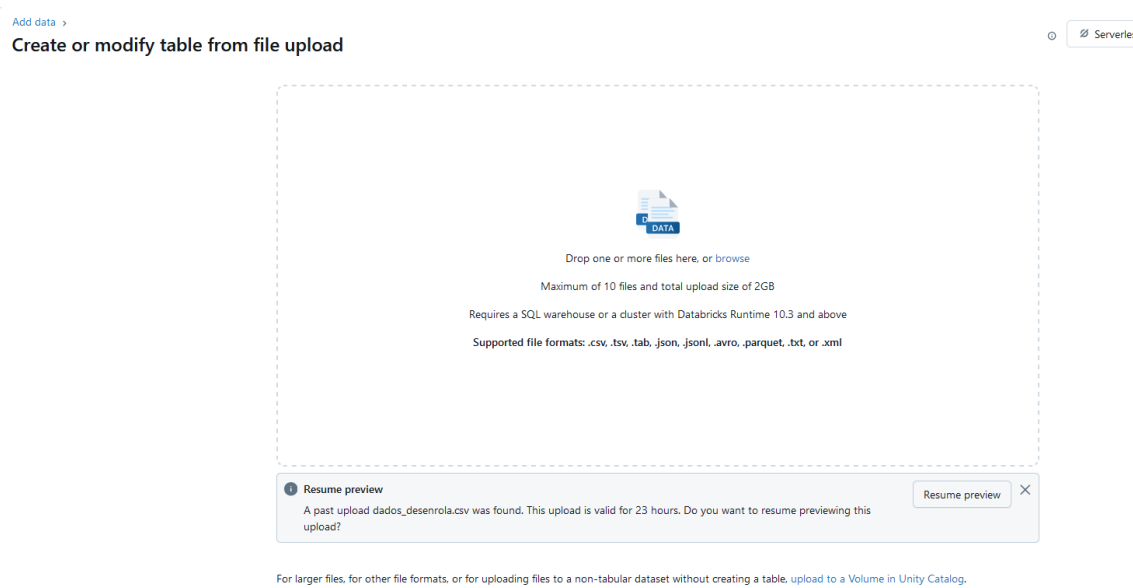
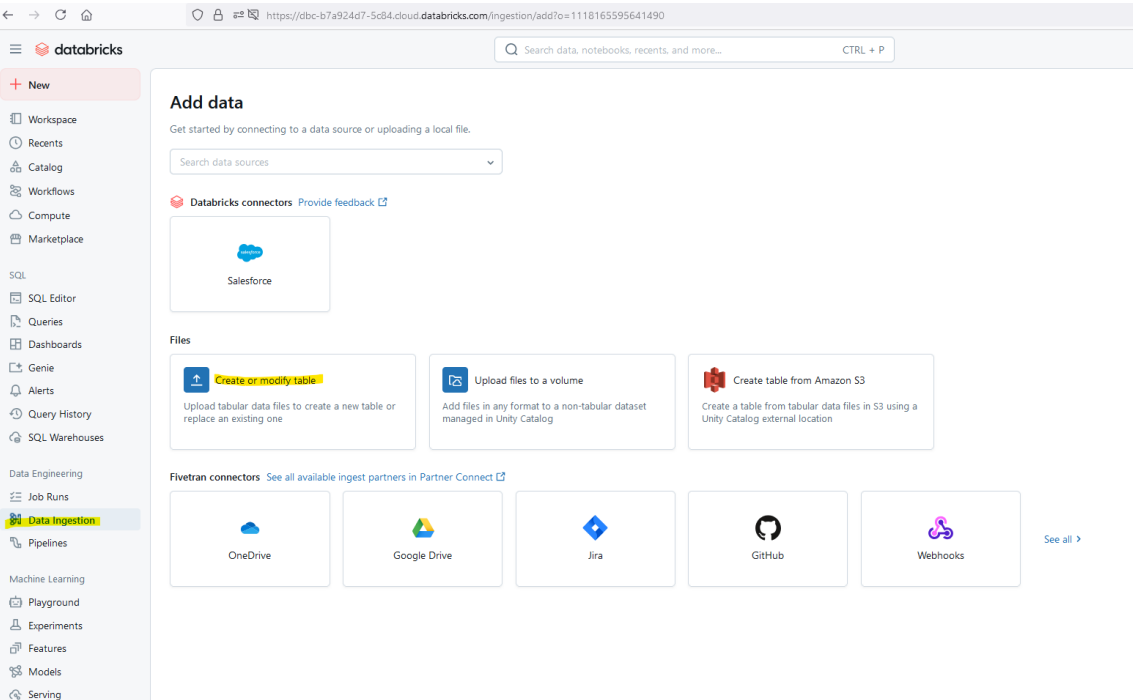
Statement	Started At	Duration	Rows read	Bytes read	Bytes written
CREATE TABLE IF NOT EXISTS default.dados_d	Mar 31, 2025, 10:10 PM	1 x 639 ms	0	0 B	0 B
USE CATALOG 'workspace'	Mar 31, 2025, 10:10 PM	178 ms	0	0 B	0 B

The screenshot shows the Databricks Catalog Explorer interface. The left sidebar is similar to the previous one. The main area displays the 'default' schema. Below the 'Description' section, there is a table listing the contents of the schema:

Name	Owner	Created at	Popularity
dados_desenrola	henrique.mfrancis@gmail.com	Mar 31, 2025, 10:10 PM	...
department	henrique.mfrancis@gmail.com	Mar 31, 2025, 09:54 PM	...
max_distance_by_week	henrique.mfrancis@gmail.com	Mar 31, 2025, 10:25 PM	---
taxi_raw_records	henrique.mfrancis@gmail.com	Mar 31, 2025, 10:25 PM	---
total_fare_amount_by_week	henrique.mfrancis@gmail.com	Mar 31, 2025, 10:25 PM	---

b. Ingestão de dados

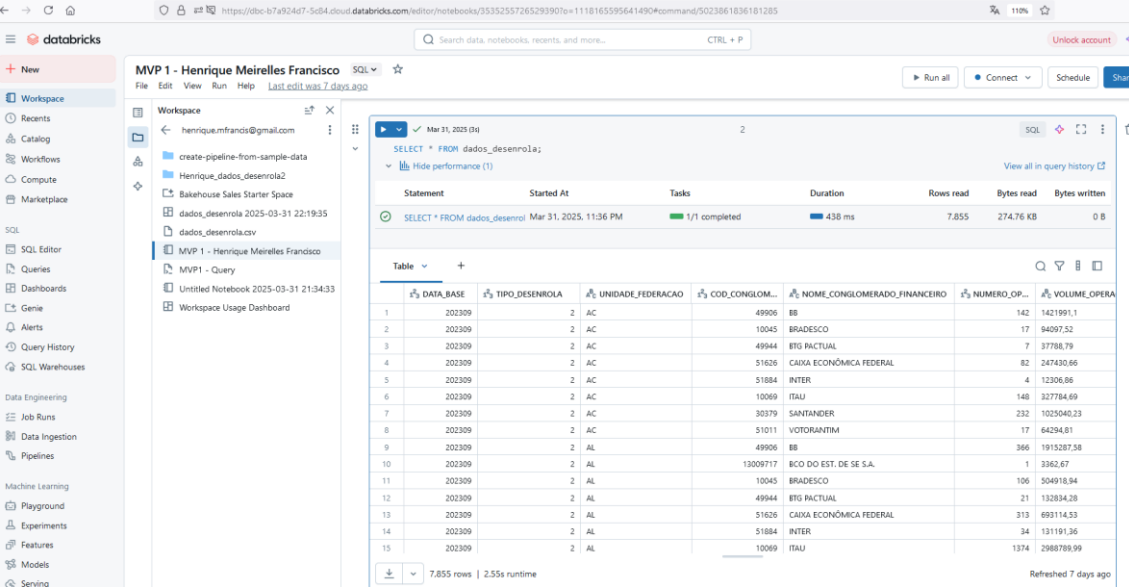
Foi utilizado a funcionalidade Data Ingestion/Create or modify table



Resultado da consulta na tabela DADOS_DESENROLA:

Notebook MVP 1 - Henrique Meirelles Francisco:

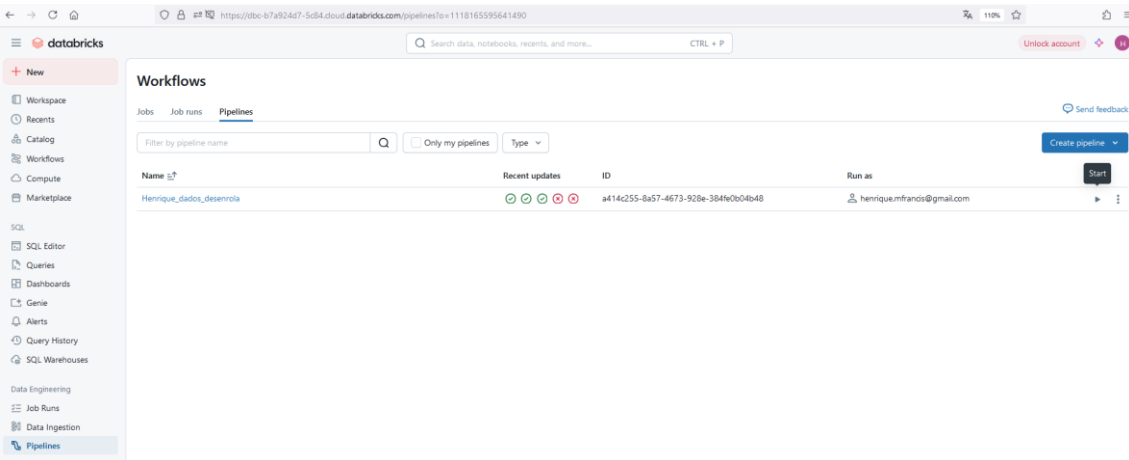
<https://dbc-b7a924d7-5c84.cloud.databricks.com/editor/notebooks/3535255726529390?o=1118165595641490>



The screenshot shows a Databricks workspace with a notebook titled "MVP 1 - Henrique Meirelles Francisco". The notebook contains a SQL query: `SELECT * FROM dados_desenrola;`. The query has been executed, and the results are displayed in a table. The table has 15 rows and 8 columns. The columns are: DATA_BASE, TIPO_DESENROLA, UNIDADE_FEDERACAO, COD_CONGLOM..., NOME_CONGLOMERADO_FINANCEIRO, NUMERO_OP..., and VOLUME_OPERA. The table shows data for various financial institutions and their operations.

	DATA_BASE	TIPO_DESENROLA	UNIDADE_FEDERACAO	COD_CONGLOM...	NOME_CONGLOMERADO_FINANCEIRO	NUMERO_OP...	VOLUME_OPERA
1	202309	2	AC	49906	BB	142	1421991.1
2	202309	2	AC	10045	BRADESCO	17	54097.52
3	202309	2	AC	49944	BTG PACTUAL	7	37788.79
4	202309	2	AC	51636	CAIXA ECONOMICA FEDERAL	82	247430.66
5	202309	2	AC	51884	INTER	4	12306.86
6	202309	2	AC	10069	ITAU	148	327784.69
7	202309	2	AC	30379	SANTANDER	232	1025040.23
8	202309	2	AL	51011	VOTORANTIM	17	64294.81
9	202309	2	AL	49906	BB	366	1915287.58
10	202309	2	AL	13009717	BCO DO EST. DE SE S.A.	1	3362.67
11	202309	2	AL	10045	BRADESCO	106	304918.94
12	202309	2	AL	49944	BTG PACTUAL	21	132834.28
13	202309	2	AL	51626	CAIXA ECONOMICA FEDERAL	313	693114.53
14	202309	2	AL	51884	INTER	34	131191.36
15	202309	2	AL	10069	ITAU	1374	2988789.99

Foi feita uma tentativa de criação de um pipeline. Mas por falta de conhecimento na plataforma DataBricks e devido à expiração do uso gratuito, não foi possível continuar com a criação do pipeline.



The screenshot shows the Databricks Pipelines page. It displays a list of pipelines. The first pipeline is named "Henrique_dados_desenrola". It has a status of "Completed" and a recent update time of "2023-09-14 14:25:55". The pipeline is owned by "henrique.mfrancis@gmail.com".

Name	Recent updates	ID	Run as
Henrique_dados_desenrola	2023-09-14 14:25:55	a414c255-8a57-4673-928e-3849c0b04b48	henrique.mfrancis@gmail.com

Screenshot of the Databricks interface showing a successful pipeline update for 'Henrique_dados_desenrola'.

Workflow: Henrique_dados_desenrola

Status: Completed (31/03/2025, 23:05:19)

Graph: The workflow graph shows a 'Streaming table' node 'taxi_raw_records' (Completed - 4s) feeding into two 'Materialized view' nodes: 'max_distance_by_...' (Completed - 4s) and 'total_fare_amount...' (Completed - 4s).

Pipeline details:

- Pipeline ID: a414c255-8a57-4673-928e-384fe0b04b48
- Pipeline type: ETL pipeline
- Source code: /Users/henrique.mfrancis@gmail.com/create-pipeline-from-sample-data/sample-DLT-pipeline-notebook
- Run as: henrique.mfrancis@gmail.com

Event log:

Time	Event Type	Message
7 days ago	user_action	User henrique.mfrancis@gmail.com started an update.
7 days ago	create_update	Update 6983d6 started by API_CALL.
7 days ago	update_progress	Update 6983d6 is WAITING_FOR_RESOURCES.
7 days ago	update_progress	Update 6983d6 is INITIALIZING.
7 days ago	update_progress	Update 6983d6 is SETTING_UP_TABLES.
7 days ago	flow_definition	Flow 'workspace':default::taxi_raw_records defined as APPEND.

Screenshot of the Databricks interface showing a failed pipeline update for 'Henrique_dados_desenrola'.

Workflow: Henrique_dados_desenrola

Status: Failed (02/04/2025, 10:23:15)

Graph: The workflow graph shows a sequence of steps: 1. Creating update (Failed), 2. Waiting for resources, 3. Initializing, 4. Setting up tables, 5. Rendering graph.

Pipeline details:

- Pipeline ID: a414c255-8a57-4673-928e-384fe0b04b48
- Pipeline type: ETL pipeline
- Source code: /Users/henrique.mfrancis@gmail.com/create-pipeline-from-sample-data/sample-DLT-pipeline-notebook
- Run as: henrique.mfrancis@gmail.com

Event log:

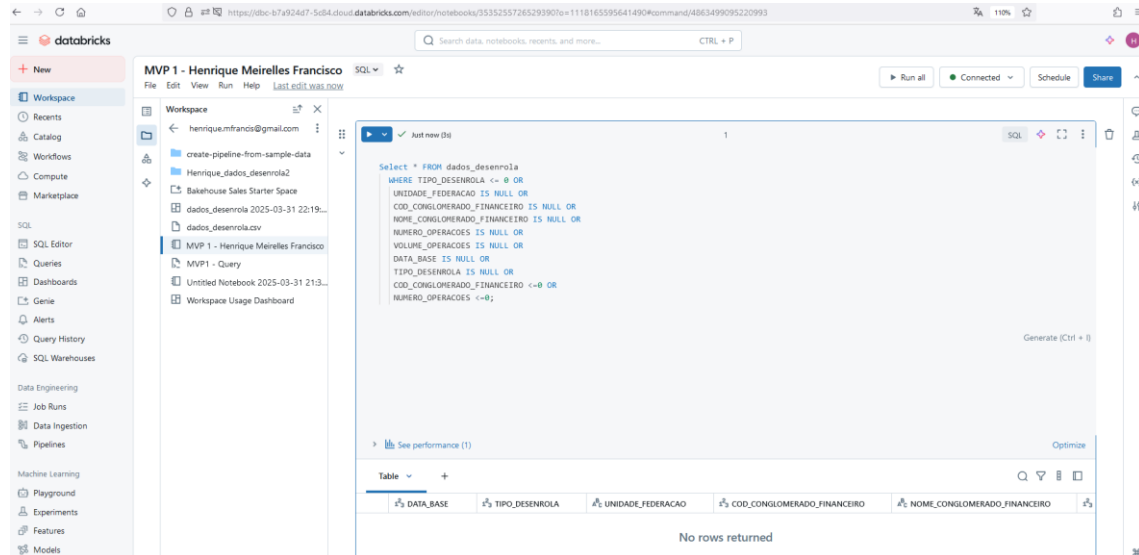
Time	Event Type	Message
6 days ago	user_action	User henrique.mfrancis@gmail.com started an update.
6 days ago	create_update	Update 38b6d2 started by API_CALL.
6 days ago	update_progress	Update 38b6d2 is FAILED.

5. Análise

a. Qualidade de dados

Para uma primeira averiguação da qualidade dos dados utilizou-se a query abaixo com o objetivo de identificar valores NULOS e valores numérico negativos ou iguais a zero.

O resultado não retornou nenhum valor o que garante a qualidade no critério acima.



The screenshot shows the Databricks SQL Editor interface. The query being executed is:

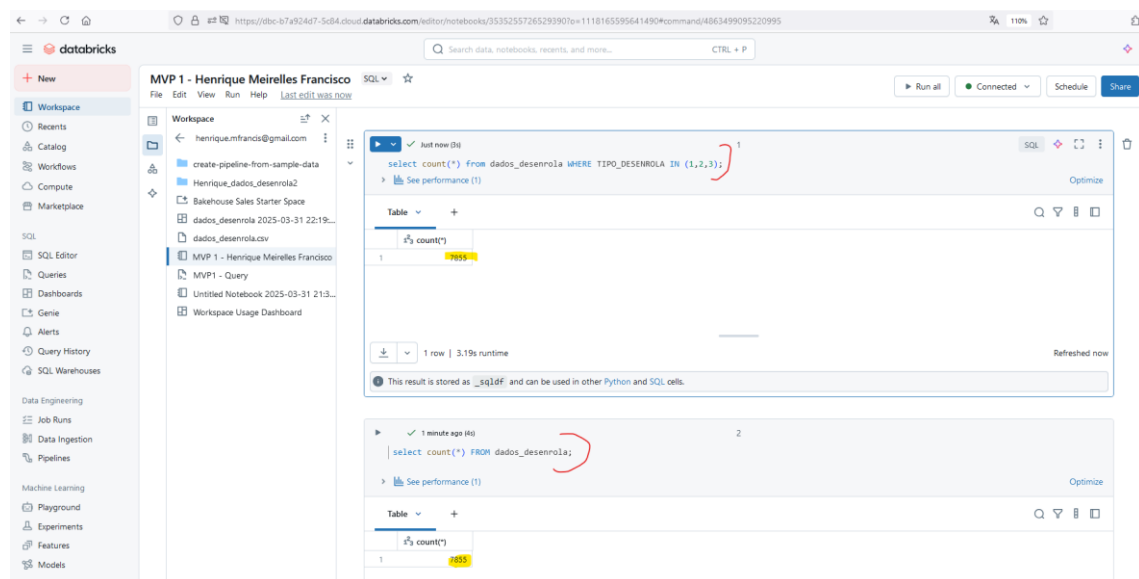
```
SELECT * FROM dados_desenrola
WHERE TIPO_DESENROLA <= 0 OR
UNIDADE_FEDERACAO IS NULL OR
COD_CONGLOMERADO_FINANCEIRO IS NULL OR
NOME_CONGLOMERADO_FINANCEIRO IS NULL OR
NUMERO_OPERACOES IS NULL OR
VOLUME_OPERACOES IS NULL OR
DATA_BASE IS NULL OR
TIPO_DESENROLA IS NULL OR
COD_CONGLOMERADO_FINANCEIRO <= 0 OR
NUMERO_OPERACOES <= 0;
```

The results table is empty, indicating that no rows were returned, which confirms the data quality for the specified criteria.

A segunda averiguação foi garantir que o domínio da coluna TIPO_DESENROLA contém apenas os valores: 1, 2 e 3.

Foi feita a comparação das quantidades nas queries abaixo.

Os valores foram idênticos, demonstrando a consistência dos valores da coluna TIPO_DESENROLA.



The screenshot shows two queries being executed in the Databricks SQL Editor to verify the domain of the TIPO_DESENROLA column.

Query 1:

```
SELECT COUNT(*) FROM dados_desenrola WHERE TIPO_DESENROLA IN (1,2,3);
```

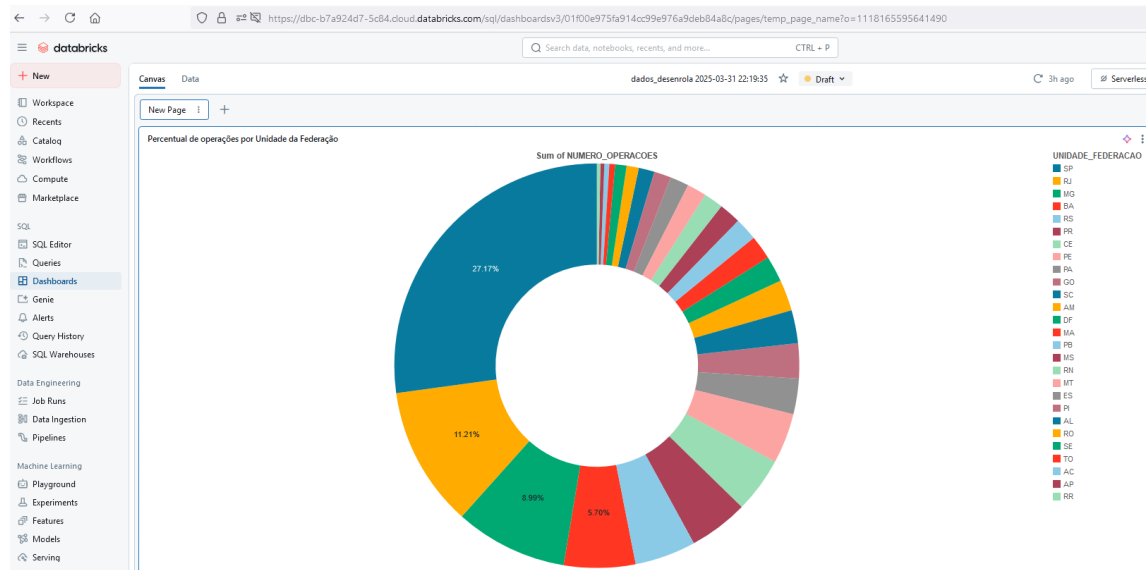
Query 2:

```
SELECT COUNT(*) FROM dados_desenrola;
```

Both queries returned a count of 1, indicating that the domain of the TIPO_DESENROLA column is consistent with the expected values (1, 2, and 3).

b. Solução do problema

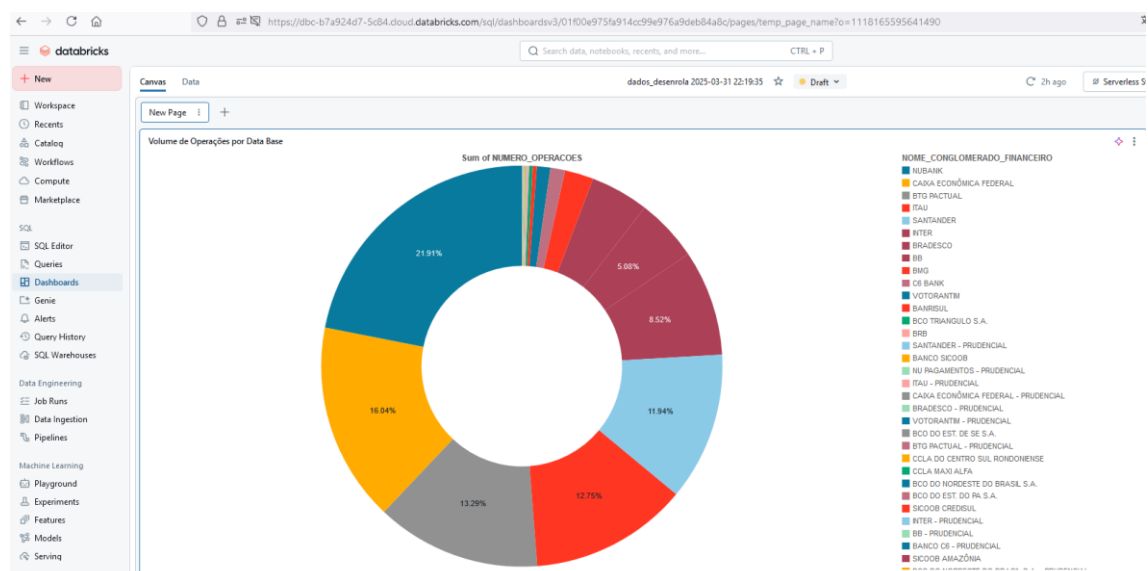
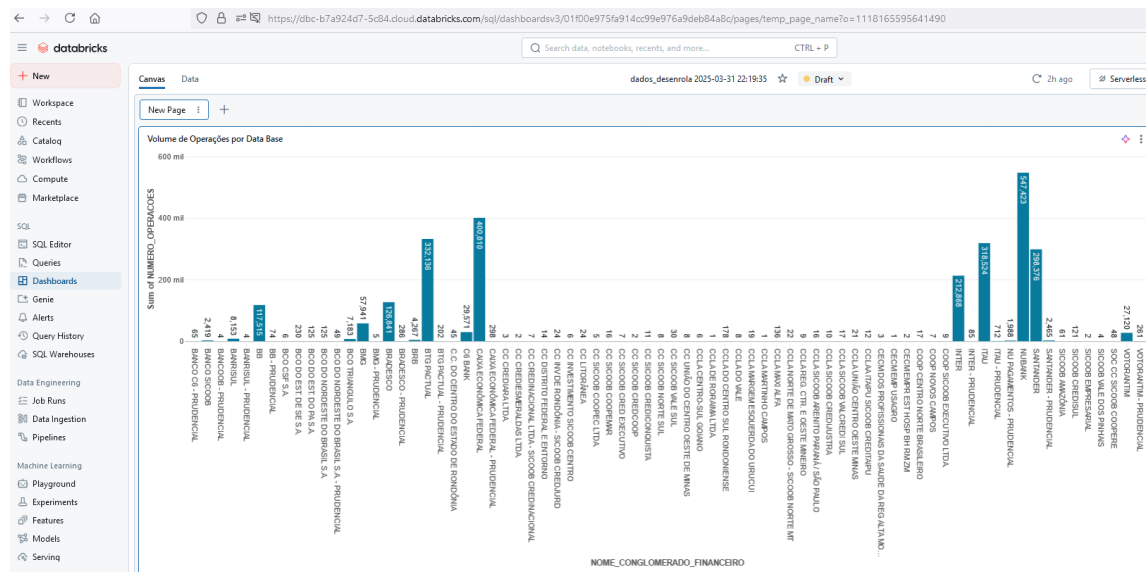
b.1- Percentual de operações por Unidade da Federação



As unidades da Federação com maior volume de operações no desenrola brasil foram SP com 27.17% das operações, RJ com 11.21% das operações e MG com 8.99% das operações.

Não é de surpreender, pois os três estados são os mais populosos do Brasil.

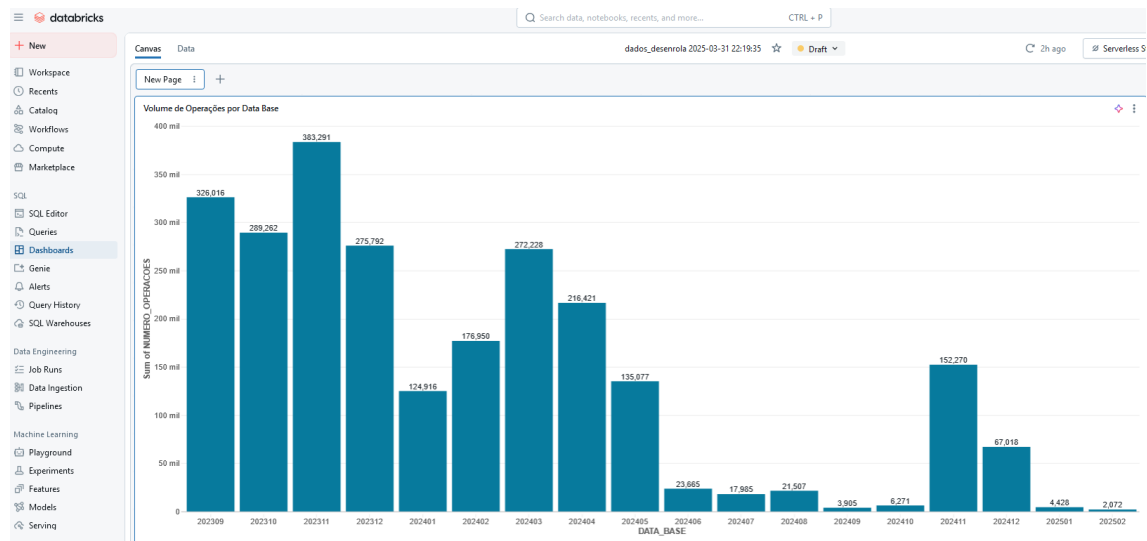
b.2 – Volume e percentual de Operações por Conglomerado Financeiro



O NUBANK foi o banco com maior volume de operações com 21.91% das operações, que foi uma surpresa por se tratar de um banco 100% digital.

A CAIXA ECONÔMICA FEDERAL teve o segundo maior volume de operações com 16.04% das operações, que por se tratar de um baco público com grande presença no Brasil, era esperado ser o primeiro em volume.

b.3 – Volume de operações por data base



Como esperado, no início do programa, 2023, tivemos um maior volume de operações e um decréscimo nos anos seguintes.

A análise revelou que São Paulo, Rio de Janeiro e Minas Gerais lideram em volume de operações, enquanto o Nubank se destacou como o conglomerado financeiro com maior volume de operações, seguido pela Caixa Econômica Federal.

Autoavaliação:

Consegui responder as perguntas feitas no objetivo. A base de dados escolhida foi bem simples contendo uma única tabela, pois meu principal objetivo foi entender todo o processo de engenharia de dados, de forma prática.

Tive bastante dificuldade no uso da plataforma databricks, principalmente por não conhecer essa plataforma. Apreendi a usá-la estudando o manual, vídeos no youtube e na tentativa e erro. Acabei por utilizar todo o crédito da licença gratuita, o que dificultou o aprendizado. No caso da geração de um pipeline não tive sucesso. Creio que com mais estudo/dicas conseguiria fazer o pipeline.

Espero poder evoluir nas demais sprints e investir numa assinatura de uma plataforma em nuvem.

URLs da plataforma com minhas atividades:

Workspace Databricks:

<https://dbc-b7a924d7-5c84.cloud.databricks.com/browse/folders/workspace?o=1118165595641490>

Notebook:

<https://dbc-b7a924d7-5c84.cloud.databricks.com/editor/notebooks/3535255726529390?o=1118165595641490>

Dashboard:

<https://dbc-b7a924d7-5c84.cloud.databricks.com/sql/dashboards?o=1118165595641490>

GIT: https://github.com/HenriqueMFrancisco/PUCRJ_P-S/blob/main/mvp1_henrique.docx