Logotipo

O conteúdo gerado por IA pode estar incorreto.

**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>

Interface gráfica do usuário

O conteúdo gerado por IA pode estar incorreto.

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>

Tela de computador com texto preto sobre fundo branco

O conteúdo gerado por IA pode estar incorreto.

***2. Coleta***

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

*Diretório local:*

Interface gráfica do usuário, Texto, Aplicativo

O conteúdo gerado por IA pode estar incorreto.

*Workspace no databricks:*

Interface gráfica do usuário, Texto

O conteúdo gerado por IA pode estar incorreto.

***3. Modelagem***

O modelo contém uma única entidade:

Dados Desenrola

Interface gráfica do usuário, Texto

O conteúdo gerado por IA pode estar incorreto.

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>

Interface gráfica do usuário, Texto, Aplicativo, Email

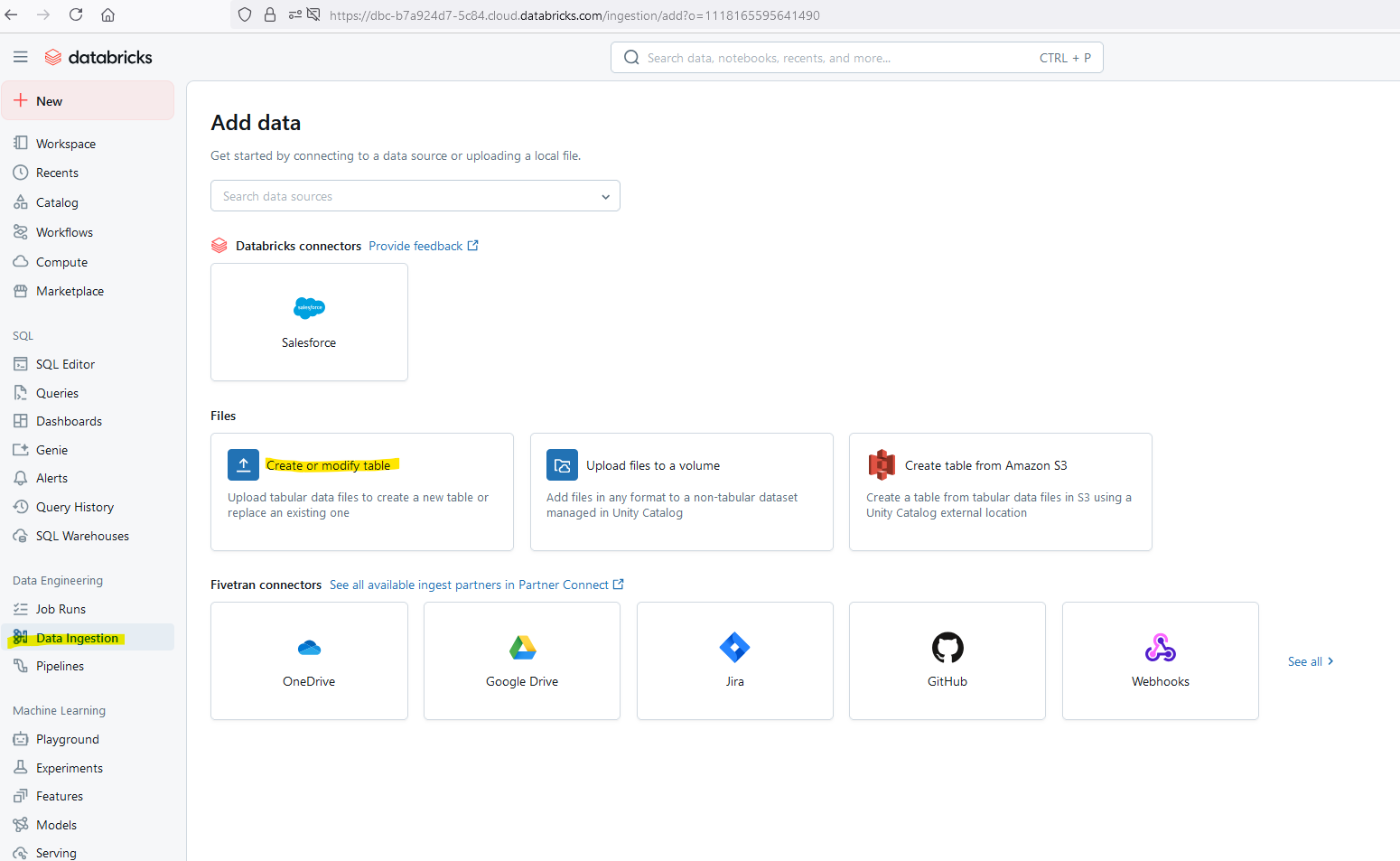
O conteúdo gerado por IA pode estar incorreto.

Tela de computador com texto preto sobre fundo branco

O conteúdo gerado por IA pode estar incorreto.

b. Ingestão de dados

Foi utilizado a funcionalidade Data Ingestion/Create or modify table



Interface gráfica do usuário, Texto, Aplicativo

O conteúdo gerado por IA pode estar incorreto.

Interface gráfica do usuário, Aplicativo

O conteúdo gerado por IA pode estar incorreto.

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>

Interface gráfica do usuário, Tabela

O conteúdo gerado por IA pode estar incorreto.

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.

Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

Texto

O conteúdo gerado por IA pode estar incorreto.

Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

**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.

Interface gráfica do usuário, Texto, Aplicativo

O conteúdo gerado por IA pode estar incorreto.

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 valore foram idênticos, demostrando a consistência dos valores da coluna TIPO\_DESENROLA.

Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

*b. Solução do problema*

*b.1- Percentual de operações por Unidade da Federação*

Gráfico, Gráfico de explosão solar

O conteúdo gerado por IA pode estar incorreto.

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*

Uma imagem contendo Gráfico

O conteúdo gerado por IA pode estar incorreto.

Gráfico, Gráfico de explosão solar

O conteúdo gerado por IA pode estar incorreto.

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*

Gráfico, Gráfico de barras

O conteúdo gerado por IA pode estar incorreto.

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. Aprendi 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>