**MVP 2 – Engenharia de Dados**

**Aluno: Vanderson Lopes Felix da Silveira**

**Painel de Obras Públicas do Brasil**

Data: 12/07/2024

# Notebook Databricks

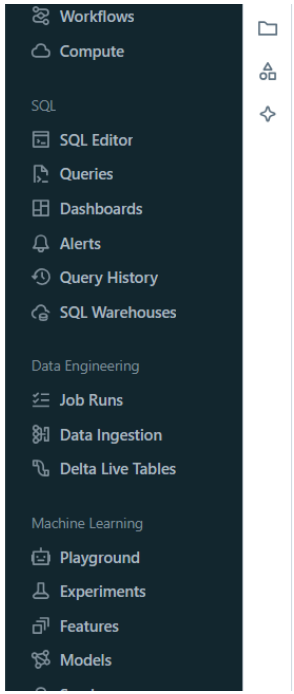
# Consultas SQL

# Passo a passo do MVP

Continua...

A tela inteira evidenciando que o código no meu notebook do Databricks (ver meu login no canto superior direito).


Continua...




Links:

- Painel de Obras: https://clusterqap2.economia.gov.br/extensions/painel-obras/painel-obras.html
- Censo 2022 IBGE: https://censo2022.ibge.gov.br/panorama/mapas.html?localidade=&recorte=N3

O Painel de Obras é um portal mantido e atualizado pelo Ministério da Gestão e da Inovação em Serviços Públicos e reúne informações de obras por todo o país e, através dele, é possível visualizar os valores investidos, a situação atual, a execução física e a execução financeira das obras.

Os dados foram coletados no dia 26/06/2024. A base está organizada em uma grande tabela em que as linhas são as obras e as colunas trazem informações como identificador da obra, órgão executor, data de início e data fim, UF, município, situação atual da obra, etc. O Catálogo de Dados constante no GitHub traz o detalhamento dessas informações.

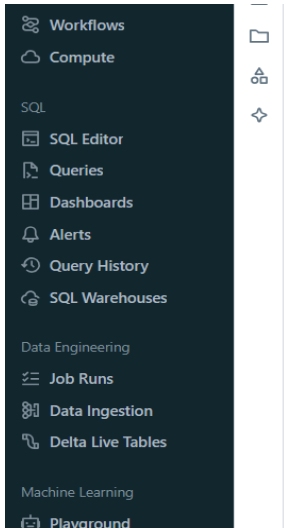
## 2. Justificativa:

O Brasil é um país em desenvolvimento e tal condição pode ser traduzida como um país que precisa atender às demandas de sua população nas suas necessidades básicas, constitucionais e estratégicas. Muitas dessas demandas são atendidas através de investimentos em estruturas e obras públicas.

Contudo, esses investimentos precisam ser feitos com eficiência, que não pode ser confundida com racionamento. Em vez disso, eficiência deve traduzir racionalidade.

O valor do investimento das obras não será objeto. O que se procura é fazer o acompanhamento dessas obras, através de consutas às base de dados. O passo-a-passo desse acompanhamento são os objetivos específicos do presente trabalho.

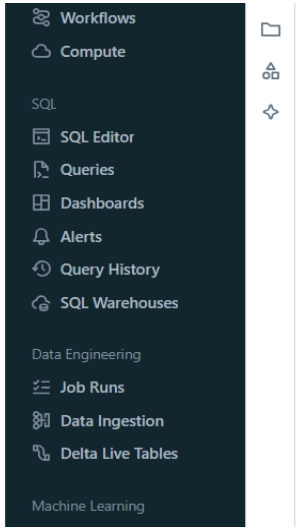
Continua...




## 3. Objetivos Específicos:

Os objetivos específicos são as perguntas que se deseja responder ao consultar a base de dados e comporão a seção de análise deste trabalho. São elas (doze ao todo):

- Qual a quantidade total de obras?
- Qual o total de investimentos em obras?
- Qual o custo médio das obras?
- Quais são as obras mais caras?
- Quais são as obras mais baratas?
- Qual estado (UF) recebeu a maior quantidade de obras?
- Qual estado (UF) recebeu o maior volume de investimentos em obras?
- Quantas obras e qual o total de investimentos por estado (UF)?
- Qual o percentual de obras em execução, paralisadas e canceladas?
- Qual estado (UF) possui a maior quantidade de obras paralisadas e de obras canceladas?
- Qual estado (UF) possui o maior volume de investimentos em obras paralisadas e de obras canceladas?
- Qual estado (UF) recebeu o maior volume de investimentos por habitante?




## 4. Desenvolvimento:

Este trabalho será desenvolvido em 4 etapas: busca de dados, coleta de dados, modelagem de dados e análise de dados.

**4.1. Busca de Dados:**

O principal critério para a escolha do banco de dados a ser utilizado neste trabalho foi a aplicação prática para a coletividade brasileira.

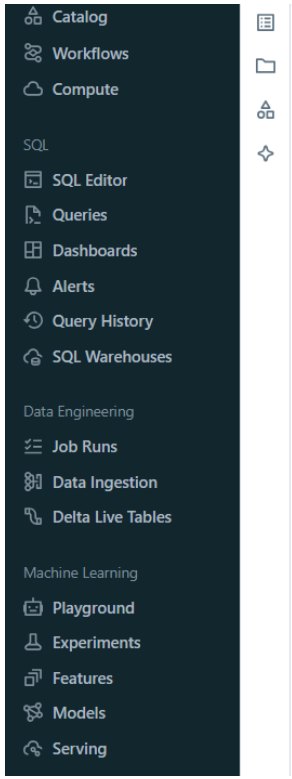
Conforme já informado na seção "Objetivo", nosso trabalho utilizou dois bancos de dados. O primeiro é o Painel de Obras do governo federal. O segundo são os dados oficiais do censo 2022 do IBGE.

O que se encontra disponível no Painel de Obras faz parte do orçamento do governo federal empenhado ao longo dos anos para a realização de obras espalhadas nos milhares de municípios das 27 unidades da federação. Por sua vez, do censo 2022 do IBGE, foi utilizada a informação sobre o tamanho da população brasileira por unidade da federação.

Por fluidez do trabalho, convém repetir nesta seção o link das duas bases de dados. Vamos a eles:

Painel de Obras: https://clusterqap2.economia.gov.br/extensions/painel-obras/painel-obras.html

Censo 2022 IBGE: https://censo2022.ibge.gov.br/panorama/mapas.html?localidade=&recorte=N3




**4.2. Coleta, Modelagem e Carga de Dados:**

4.2.1. Coleta da Dados:

O processo de coleta das informações compreende atividades de ETL (Extração, Transformação e Carga), que não ocorreu de forma linear (pontual) no desenvovimento deste trabalho. Ou seja, em alguns momentos, a atividade de transformação, por exemplo, foi necessária quando a consulta pelo SQL evidenciava alguma inconsistência nos bancos de dados.

A atividade de extração consistiu em fazer o download dos arquivos do Painel de Obras no site do Portal da Transparência (foram 27 arquivos, um para cada estado) e do arquivo do censo 2022 do site do IBGE em formato .xlsx (Excel) para o computador pessoal.

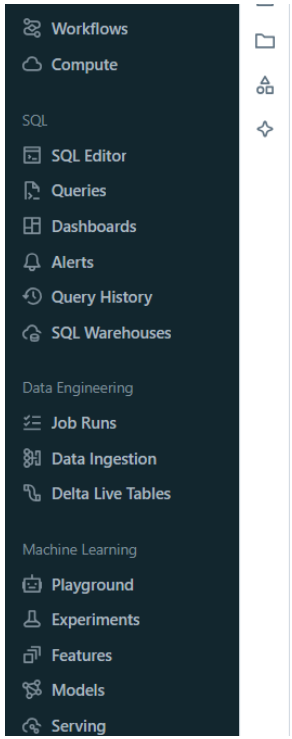
A base de dados do Painel de Obras é robusta, contendo mais de 185 mil linhas (obras) 42 colunas, de modo que sua importação de uma só vez não foi possível. Como alternativa, essa base original foi desmembrada por estado (unidade da federação - UF) e cada um dos 27 arquivos foi exportado em formato Excel (.xlsx) para um computador pessoal.

Como o DataBricks, plataforma usada para carga e armazenamento dos dados, não é compatível com arquivos no formato Excel (pelo menos, não se coseguiu fazer a importação), foi necessário fazer a primeira transformação, modificando o formato dos arquivos: do .xlsx para o .csv, além da remoção de caracteres especiais nos nomes das colunas como cedilhas, acento, til, etc. Essa transformação foi feita em Python de acordo com o código no collab incluído no diretório do GitHub.

Em seguida, esses 27 arquivos transformados foram importados para o Catálogo do DataBricks e inseridos em 3 arquivos reunindo informações de 9 estados (UF) cada: “painel_1”, “painel_2” e “painel_3”.

Já a base da dados do IBGE foi exportada do site do IBGE contendo 27 linhas (uma para cada UF) e duas colunas ("Unidades da Federação e "pessoas"). Este arquivo original foi importado, já em .csv, para um computador pessoal e passou pela seguinte transformação: a coluna "unidades da Federação" foi substituída pela coluna "UF", onde o conteúdo das linhas com os nomes dos estados por extenso foi substituído pelas respectivas siglas. Esse arquivo foi nomeado de “censo_2022_uf” e também foi importado para o Catálogo do DataBricks.

Continua...




4.2.2. Carga de Dados:

Portanto, quatro bases de dados foram carregadas no Catálogo da Plataforma do DataBricks: "painel_1", "painel_2", "painel_3" e "censo_2022_uf".

Importante ressaltar que todos os arquivos permanecem na base de dados da plataforma. A persistência dos dados está evidenciada na Figura 1 do arquivo "ANEXO I_Evidências Sprint 2" no link do diretório do GitHub.

Antes de prosseguir, comvém explicar um pouco mais sobre essa plataforma de dados. Afinal, o que é a Plataforma DataBricks?

O site da ALURA (https://www.alura.com.br/artigos/databricks-o-que-e-para-que-serve), traz a seguinte explicação sobre a Plataforma Databricks (texto adaptado):

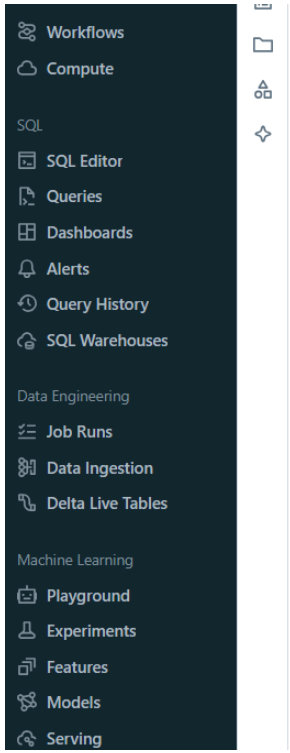
A plataforma Databricks como uma solução de computação em nuvem que pode ser usada para processamento, transformação e exploração de grandes volumes de dados. Ela foi projetada para permitir que os usuários se concentrem em análises de dados avançadas e na tomada de decisões baseadas em dados, de uma forma mais simples. A plataforma é altamente escalável e pode ser configurada para trabalhar com vários serviços em nuvem, incluindo Amazon Web Services (AWS).

O Databricks utiliza clusters para o processamento de grandes volumes de dados de forma distribuída, o que a torna uma ferramenta eficiente e escalável. Ela possibilita ao próprio usuário utilizar sua interface simples para gerenciar seus clusters.

Para usar a plataforma Databricks, os usuários criam notebooks, que são documentos interativos que permitem escrever e executar código para processar dados. Os notebooks podem incluir código em várias linguagens de programação, como Python, R e SQL.

Voltando para a coleta de dados propriamente dita, feita a inserção das bases de dados, os arquivos "painel_1", "painel_2" e "painel_3" foram consolidados em apenas uma base, chamada "painel", restaurando, na verdade, o conteúdo original da base de dados do site Painel de Obras em apenas um arquivo. A mesma rotina (em Python) que executa essa transformação também substitui o espaço pelo caracter "underscore" nos nomes das colunas. Essa rotina segue abaixo.

Continua...



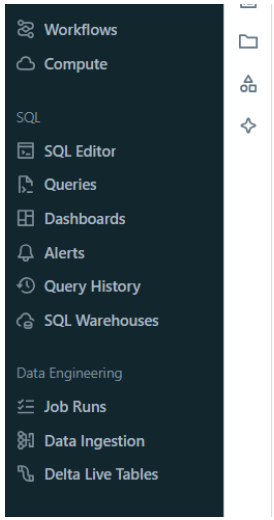

```
%python

#Consolida os 3 arquivos e remove os espaços do cabeçalho da tabela resultado da união

df = spark.sql("SELECT * FROM (SELECT * FROM painel_1 UNION SELECT * FROM painel_2 UNION SELECT * FROM painel_3)")

for column in df.columns: #lista com os nomes das colunas
    df=df.withColumnRenamed(column,column.replace(" ","_"))
df.write.mode("overwrite").saveAsTable("painel") #sobrescreve
```

A query abaixo exibe as 5 primeiras linhas da base da dados "painel".

```
SELECT *
FROM painel LIMIT(5)
```

Além disso, foram necessárias algumas transformações para limpar a base de dados. Por exemplo, havia linhas de "painel" em que o conteúdo das colunas "ID_obra" e "Investimento_Total" era nulo (is null) ou zero. As duas rotinas SQL (consulta) que seguem exemplificam essas inconsistências.

Continua...



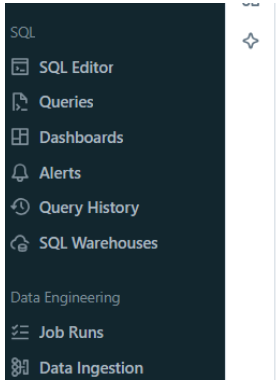

10

```
--IDENTIFICANDO AS LINHAS ONDE O ID_OBRA É NULO (exibe as 5 primeiras):

SELECT *
FROM painel
WHERE ID_Obra is NULL LIMIT (5)
```

11

```
--IDENTIFICANDO AS LINHAS ONDE O INVESTIMENTO_TOTAL É NULO (exibe as 5 primeiras):

SELECT *
FROM painel
WHERE Investimento_Total is NULL LIMIT (5)
```

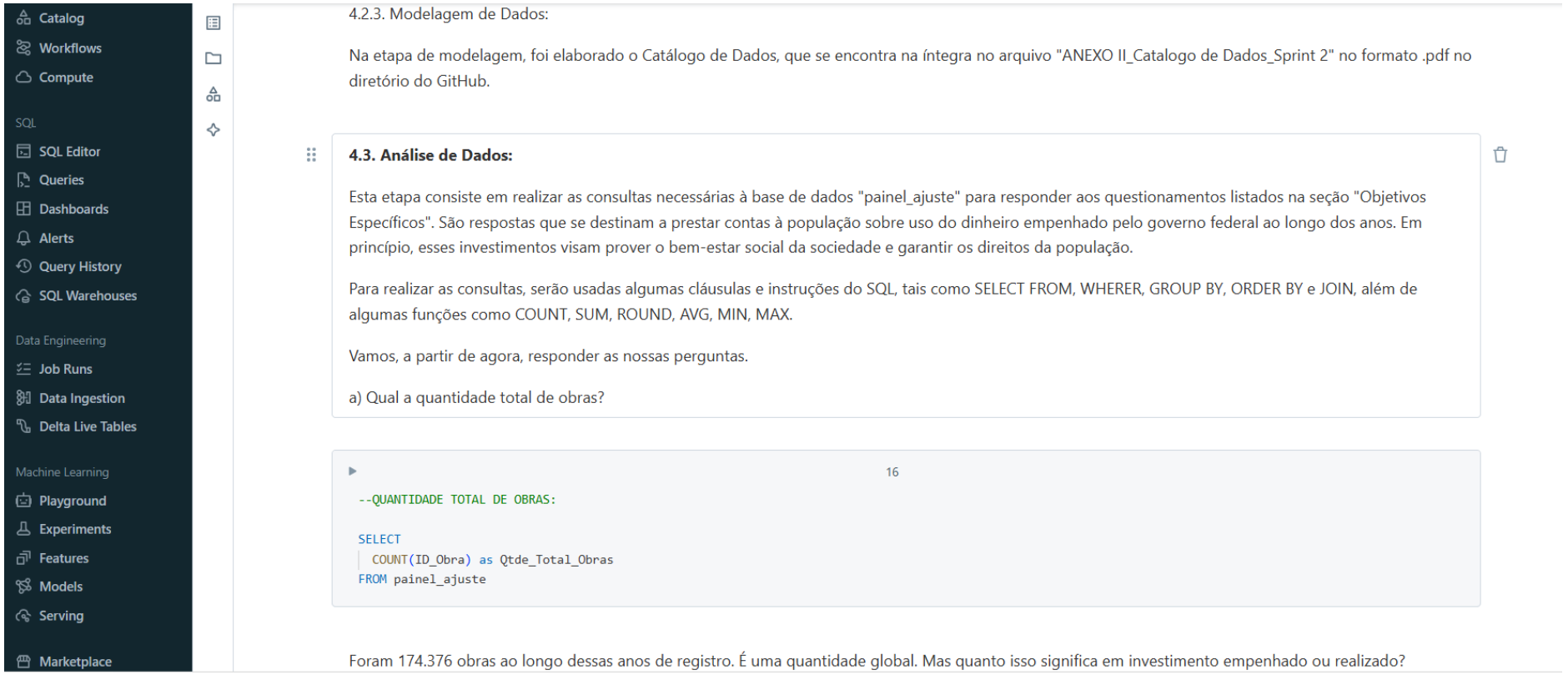
Para corrigir essas e outras inconsistências, essas linhas foram removidas pela rotina abaixo (em Python). A Tabela resultante é a "painel_ajuste".



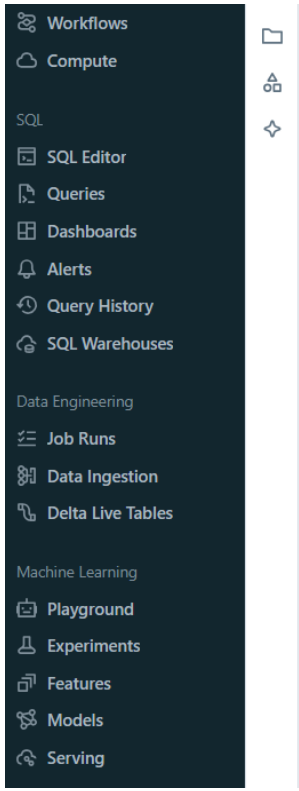

13

```
%python

#LIMPEZA DE DADOS (exclusão de algumas linhas)
import pyspark.sql.functions as F

df = spark.sql('select * from painel where ID_Obra is not null and ID_Obra != 0 and Investimento_Total != 0') #Manteve apenas ID_obra não nulo, ID_obra
diferente de zero e investimento total diferente de zero

df = df.where(F.length('UF')==2) #Retirou aquela repetição de caracteres das siglas dos estados na coluna UF"
df.createOrReplaceTempView("painel_ajuste")  #Cria tabela temporária a partir do painel_ajuste
```

Continua...


Catalog
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Delta Live Tables
Machine Learning
Playground
Experiments
Features
Models
Serving
Marketplace
4.2.3. Modelagem de Dados:
Na etapa de modelagem, foi elaborado o Catálogo de Dados, que se encontra na íntegra no arquivo "ANEXO II_Catalogo de Dados_Sprint 2" no formato .pdf no diretório do GitHub.
4.3. Análise de Dados:
Esta etapa consiste em realizar as consultas necessárias à base de dados "painel_ajuste" para responder aos questionamentos listados na seção "Objetivos Específicos". São respostas que se destinam a prestar contas à população sobre uso do dinheiro empenhado pelo governo federal ao longo dos anos. Em princípio, esses investimentos visam prover o bem-estar social da sociedade e garantir os direitos da população.
Para realizar as consultas, serão usadas algumas cláusulas e instruções do SQL, tais como SELECT FROM, WHERER, GROUP BY, ORDER BY e JOIN, além de algumas funções como COUNT, SUM, ROUND, AVG, MIN, MAX.
Vamos, a partir de agora, responder as nossas perguntas.
a) Qual a quantidade total de obras?
16
--QUANTIDADE TOTAL DE OBRAS:
SELECT
COUNT(ID_Obra) as Qtde_Total_Obras
FROM painel_ajuste
Foram 174.376 obras ao longo dessas anos de registro. É uma quantidade global. Mas quanto isso significa em investimento empenhado ou realizado?

Continua...




b) Qual o total de investimentos em obras?

```
18
--INVESTIMENTO TOTAL EM OBRAS (R$ milhões):

SELECT
  ROUND(SUM(Investimento_Total)/1000000,2) as Investimento_Total
FROM painel_ajuste
```

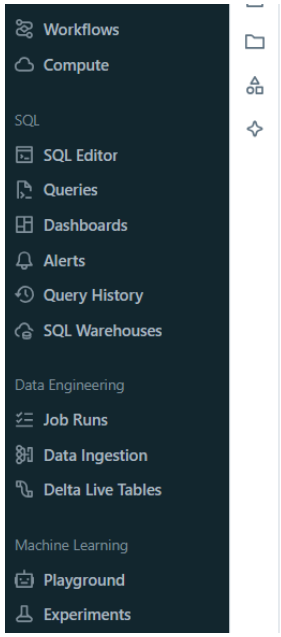
Ao longo desses anos registrados, o valor total investido foi de R$ 510,914 bilhões.

c) Qual o custo médio das obras?

```
20
--CUSTO MÉDIO DAS OBRAS (R$ milhões):

SELECT
  ROUND(AVG(Investimento_Total)/1000000,2) as Custo_Medio
FROM painel_ajuste
```

O custo médio é de R$ 2,93 milhões por obra.

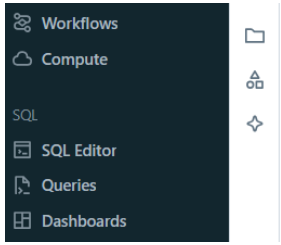
Continua...




d) Quais são as obras mais caras?

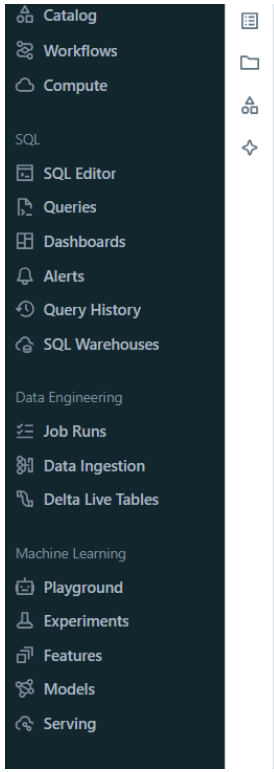
```
23
--INVESTIMENTO MAIS CARO (R$ milhões):

SELECT
  round(MAX(Investimento_Total)/1000000,2)
FROM painel_ajuste
```

```
24
--INFORMAR A OBRA MAIS CARA:

SELECT *
FROM painel_ajuste
WHERE Investimento_Total = (SELECT MAX(Investimento_Total) FROM painel)
```




```
25
--ORDENAR AS OBRAS EM ORDEM DECRESCENTE DE INVESTIMENTO:

SELECT *
FROM painel_ajuste
ORDER BY Investimento_Total DESC LIMIT (3)
```

Continua...




A obra mais cara custa R$ 76,9 bilhões. Trata-se da construção de uma infraestrutura hídrica, portos e hidrovia (tipo barragem) no município de Jucurutu no Rio Grande do Norte. A construção de uma barragem costuma ser, de fato, uma obra de capital intensivo. De toda forma, R$ 76 bilhões é um valor que chama atenção. Outro destaque é o fato de a obra ter iniciado em 2013 e ainda não ter terminado. Alías, sua execução física é de apenas 1%. O que ocorre com esta obra?

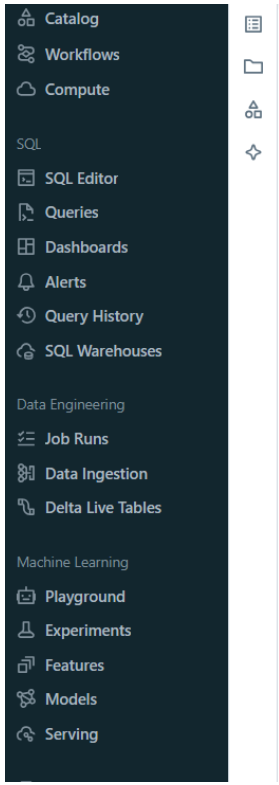
Com base nos dados constantes no Painel de Obras, a segunda obra da lista também merece acompanhamento mais próximo. Construção de de estaleiro em Itaguaí no RJ ao custo de R$ 16,39 bilhões iniciada em 2009 e ainda em execução. Já são 15 anos de obra com 80% de execução física, com previsão de terminar em 2031.

e) Quais são as obras mais baratas?

```
27
--INVESTIMENTO MAIS BARATO (R$):

SELECT
  MIN(Investimento_Total)
FROM painel_ajuste
```

```
28
--INFORMAR A OBRA MAIS BARATA:

SELECT *
FROM painel_ajuste
WHERE Investimento_Total = (SELECT MIN(Investimento_Total) FROM painel_ajuste)
```

Continua...




29

```
--ORDENAR AS OBRAS EM ORDEM CRESCENTE DE INVESTIMENTO:

SELECT *
FROM painel_ajuste
ORDER BY Investimento_Total ASC LIMIT(3)
```

Segundo a coluna Investimento_Total, a obra mais barata custou R$ 1 (um real), num provável erro de preenchimento, o que pode ser indicado pelo fato de a coluna "Valor_Desembolsado" estar preenchida com o valor de R$ 427.494,07. De fato, esse último valor parece fazer mais sentido para uma obra de um laboratório de pesquisas avançadas na Universidade Federal do Pará concluída em 2010.

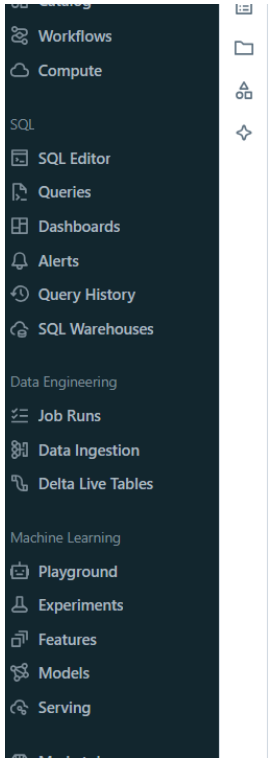
O mesmo equívoco de preenchimento parece ter ocorrido nas duas obras seguntes, ambas no Acre.

f) Qual estado (UF) recebeu a maior quantidade de obras?

32

```
--QUANTIDADE DE OBRAS POR UF (em ordem decrescente):

SELECT UF,
  COUNT(*) AS Quantidade_Total
FROM painel_ajuste
GROUP BY UF
ORDER BY Quantidade_Total DESC LIMIT (5)
```

Continua...




O Estado de São Paulo foi o que mais recebeu obras, num total de 18.396. Trata-se da unidade da federação mais populosa do país. Em seguida, vieram Minas Gerais e Rio Grande do Sul com 18.311 e 15.109, respectivamente. Interessante notar que os 5 primeiros estados concentram cerca de 43,2% do total das obras e concentram 49,9% da população brasileira.

Vamos ver se essa proporção se mantem quando se fala em valor investido?

g) Qual estado (UF) recebeu o maior volume de investimentos em obras?

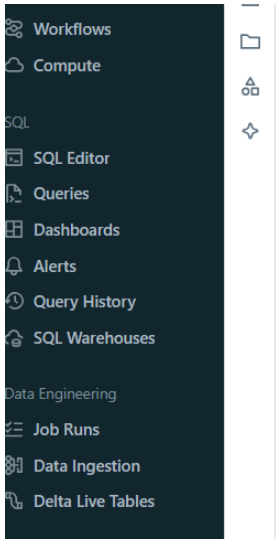
34

```
--INVESTIMENTO EM OBRAS POR UF (em ordem decrescente) (R$ milhões):

SELECT UF,
  ROUND(SUM(Investimento_Total)/1000000,2) AS Investimento_UF
FROM painel_ajuste
GROUP BY UF
ORDER BY Investimento_UF DESC LIMIT (5)
```

O Rio Grande do Norte foi quem recebeu o maior investimento ao logo desses anos: cerca de R$ 85,6 bilhões. Desses, R$ cerca de 76,9 bilhões foram do investimento na obra da barragem em Jucurutu (quase 90%).

Além disso, os cinco estados que mais receberam investimentos respondem por cerca de 52% do investimento total do Brasil ao longo desses anos, sendo que esses mesmos estados somam, juntos, 45,8%. Em cpmparação com a métrica da quantidade de obras, existe um pequeno desbalanceamento, muito provocado pela investimento "campeão" da barragem no RN.

Continua...

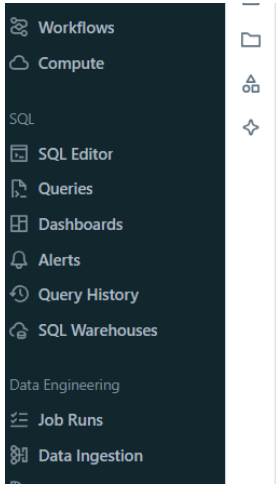



h) Quantas obras e qual o total de investimentos por estado (UF)?

37

```
--QUANTIDADE DE OBRAS E TOTAL INVESTIMENTO EM OBRAS POR UF:

SELECT UF,
  COUNT(*) AS Quantidade_Total,
  ROUND(SUM(Investimento_Total)/1000000,2) AS Investimento_Total
FROM painel_ajuste
GROUP BY UF
ORDER BY Investimento_Total ASC
```

O Distrito Federal foi a unidade da federação que recebeu a menor quantidade de obras (750) e o estado do Espírito Santo foi o que recebeu o menor volume desse tipo de investimento (R$ 4,1 bilhões).



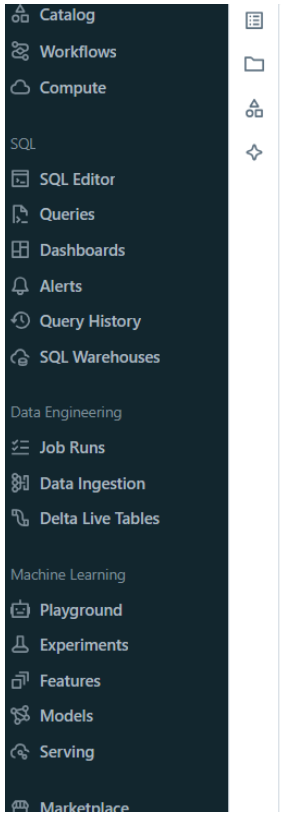

i) Qual o percentual de obras em execução, paralisadas e canceladas?

40

```
SELECT
  Situacao_Atual,
  COUNT(*) AS Qtde_Obras,
  ROUND(SUM(Investimento_Total)/1000000,2) AS Investimento_Total
FROM painel_ajuste
GROUP BY Situacao_Atual
ORDER BY Investimento_Total DESC
```

As obras que estão paralisadas correspondem a cerca de 11% do investimento total (R% 55,1 bilhões). Por sua vez, as obras que foram canceladas são 6,5% (R$ 33,6 bilhões). Já as obras concluídas somam R$ 165,8 bilhões (32,4%) enquanto as obras em execução somam R$ 233 bilhões (cerca de 45,6%).

Continua...




j) Qual estado (UF) possui a maior quantidade de obras paralisadas e de obras canceladas?

42

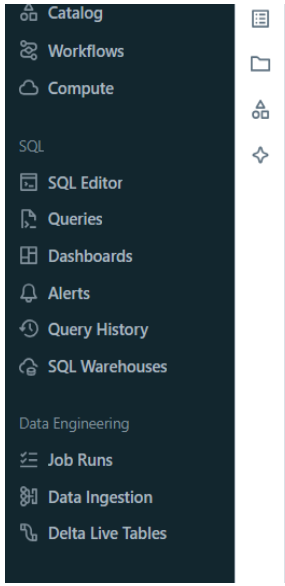
```
--QUANTIDADE DE OBRAS PARALISADAS POR UF (em ordem decrescente):

SELECT UF,
  COUNT(*) AS Obras_Paralisadas
FROM painel_ajuste
WHERE Situacao_Atual = 'Paralisada'
GROUP BY UF
ORDER BY Obras_Paralisadas DESC LIMIT (5)
```

O estado com o maior número de obras paralisadas é Minas Gerais: 264. Em seguida vem o estado do Pará (207) e São Paulo (199).

44

```
--QUANTIDADE DE OBRAS CANCELADAS POR UF (em ordem decrescente):

SELECT UF,
  COUNT(*) AS Obras_Canceladas
FROM painel_ajuste
WHERE Situacao_Atual = 'Cancelada'
GROUP BY UF
ORDER BY Obras_Canceladas DESC LIMIT (5)
```

Continua...



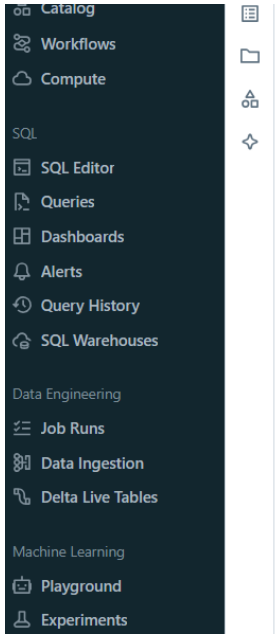

E Minas Gerais também é o campeão de obras canceladas (2628), seguido por São Paulo (2478) e Rio de Janeiro (1582).

k) Qual estado (UF) possui o maior volume de investimentos em obras paralisadas e de obras canceladas?

```
---TOTAL DE INVESTIMENTO DE OBRAS PARALISADAS POR UF (em ordem decrescente) (R$ milhões)

SELECT UF,
  ROUND(SUM(Investimento_Total)/1000000,2) AS Investimento_Paralisado
FROM painel_ajuste
WHERE Situacao_Atual = 'Paralisada'
GROUP BY UF
ORDER BY Investimento_Paralisado DESC LIMIT (5)
```

Se na quantidade de obras paralisadas, MG é o principal estado, no volume de investimentos, o topo da classificação fica com o estado de SP com aproximadamente R$ 10,6 bilhões. Isso significa que dos R$ 58,5 bilhões que foram empenhados para SP, cerca de 18% está paralisado.

Continua...




48

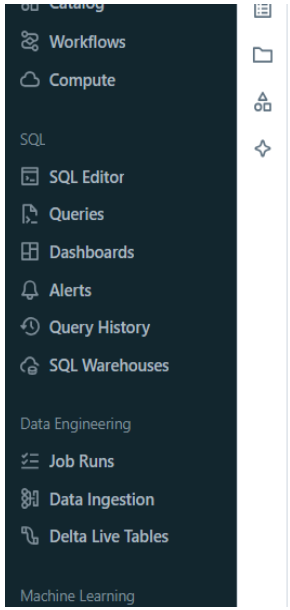
```sql
---TOTAL DE INVESTIMENTO DE OBRAS CANCELADAS POR UF (em ordem decrescente) (R$ milhões)

SELECT UF,
  ROUND(SUM(Investimento_Total)/1000000,2) AS Investimento_Cancelado
FROM painel_ajuste
WHERE Situacao_Atual = 'Cancelada'
GROUP BY UF
ORDER BY Investimento_Cancelado DESC LIMIT (5)
```

No quesito investimentos que foram cancelados, o estado de Mato Grosso teve o maior valor de R$ 8,3 bilhões. É um valor muito alto, que corresponde a cerca de 25% do total de investimentos que foram cancelados no país todo (8,3 / 33,6) (ver a tabela consulta da pergunta "i"). A query abaixo permite visualizar grande parte dessas obras foram canceladas por problemas técnicos de execução ou revisão de projeto executivo.

50

```sql
SELECT *
FROM painel_ajuste
WHERE Situacao_Atual = 'Paralisada' AND UF = 'MT'
ORDER BY Investimento_Total DESC
```

Continua...




I) Qual estado (UF) recebeu o maior volume de investimentos por habitante?

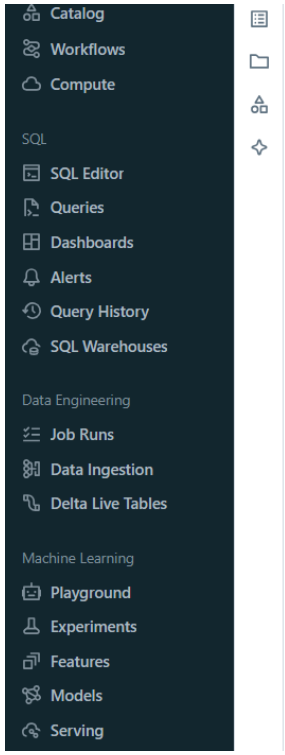
52

```
DROP TABLE IF EXISTS Obras_Estado
```

53

```
--CRIANDO TABELA - QUANTIDADE DE OBRAS E TOTAL INVESTIMENTO EM OBRAS POR UF:

CREATE TABLE Obras_Estado AS
SELECT UF,
  COUNT(*) AS Quantidade_Total,
  ROUND(SUM(Investimento_Total)/1000000,2) AS Investimento_Total
FROM painel_ajuste
GROUP BY UF
ORDER BY Investimento_Total DESC
```

Continua...

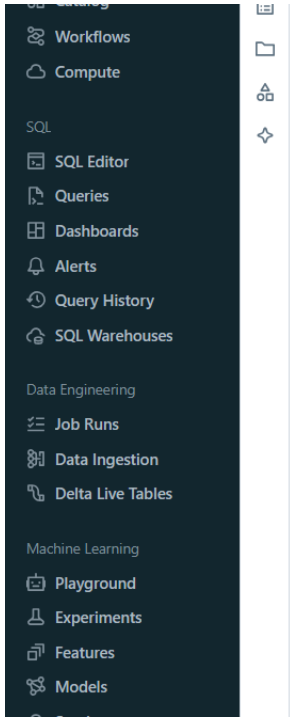



```sql
--INNER JOIN

--Faça uma consulta que tenha todas as colunas da tabela Obras_Estado e a coluna UF da tabela painel_ajuste

SELECT o.UF,
c.pessoas,
o.Quantidade_Total,
o.Investimento_Total,
ROUND(o.Investimento_Total/c.pessoas *1000000,2) as Investimento_Pessoa
FROM Obras_Estado AS o
JOIN hive_metastore.default.censo_2022_uf as c
ON o.UF = c.UF
ORDER BY Investimento_Pessoa DESC LIMIT (5)
```

Os estados do Rio Grande do Norte e Roraima destoam na relação de Investimento em Obras por habitante. O primeiro tem uma relação de R$ 25.910 / habitante; o segundo, R$ 16.363 / habitante. O terceiro colocado é o estado do Amapá com cerca de R$ 6.466 / habitante.

Ao final desta etapa de análise dos dados, conseguiu-se elaborar consultas via SQL e responder todas as 12 perguntas elaboradas na seção Objetivos específicos. São respostas que conseguem traçar um mapa, ainda que não exaustivo, das preocupações do gestor público. Afinal, as obras que foram canceladas eram mesmo necessárias? Qual o escopo da obra de R$ 76,9 bilhões em Jurucutu (RN)?

Continua...




## 5. Autoavaliação:

A Engenharia de Dados (seus conceitos e técnicas) não é um fim e si mesma. É uma ferramenta de suporte à decisão. Sua aplicação na vida pública indica trilhas de melhoria no acompanhamento de projetos.

Para este trabalho, vislumbram-se duas linhas de melhoria: a primeira diz respeito à base de dados e enriquecimento da análise obtida; a segunda melhoria seria no processo de coleta e extração doa dados.

A primeira delas é a possibilidade futura de associar esses dados com a vida prática da população como, por exemplo, o nível de emprego com carteira assinada e vínculo empregatício. Ora, toda obra pública, em seu detalhe, traz (ou deveria trazer) a quantidade de empregos previstos. Se o país já programou o dispêndio do investimento em determinada obra e essa obra foi cancelada ou paralisada, isso significa, em tese, empregos que não foram gerados. Não se trata de novos investimentos, pois esses recursos já foram empenhados (ou pelo menos programados). Em resumo, no contexto deste trabalho, se uma dessas obras paralisadas ainda for necessária, elas deveriam ser imediatamente reativadas.

Uma nova pergunta seria adicionada à lista de objetivos específicos: Qual a previsão de empregos das obras paralisada (total e por estado)?

Para respondê-la, buscar-se-ia uma base de dados que reunisse a informação sobre a quantidade de empregos previstos por cada obra.

E a resposta poderia compor uma ação governamental (política pública) para ammento de emprego formal no país e, por conseguinte, aumento de renda para a população, sobretudo para os brasileiros de menor renda.

O segundo ponto de melhoria seria na etapa de extração dos dados. No presente trabalho, a extração foi feita de forma não automatizada. Como vanço, poderia ser implementado o webscraping nos sites onde se encontram as bases de dados. A atividade de webscraping (raspagem de dados) é uma ferramenta muito usada para automatizar processos de extração, coleta e consulta de dados e informações públicas. É muito bem-vinda quando a base de dados é atualizada com frequência.