**MVP Rodolfo Montoya Disciplina: Engenharia de dados Data de entrega: 04 de Julho de 2024**

**Objetivo**

Objetivo deste MVP, é avaliar a capacidade estrutural das pontes nas estradas dos Estados Unidos, verificando se existem profissionais suficientes para realizar trabalhos de inspeção e avaliando orçamentos necessários para realizar inspeções, projetos e manutenções. Nossas perguntas que queremos responder seriam: • Risco estrutural das pontes? • Frequência necessária de inspeção? • Quantidades de oportunidades e profissionais? • Necessidade de investimento?

**Plataforma**

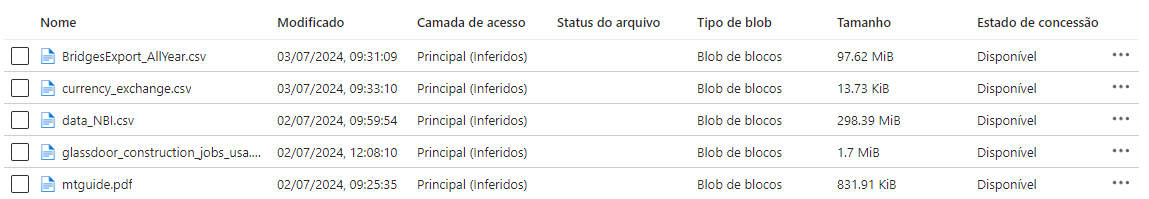
Direcionamos a Plataforma Databricks. Sendo que dentro do Microsoft Azure, temos esta ferramenta de Databricks e toda a arquitetura de dados será realizada na nuvem do Azure. Detalhamento A escolha de nossos dados foi obtida de pesquisas de informações internas, raspagem de dados do site da ASCE, classificados americanos, assim como do site kaggle. Dados utilizados: • Data.NBI.csv obtido do kaggle - <https://www.kaggle.com/datasets/broach/build-bridges-not-walls>; • mtguide.pdf, obtido do site da internet <https://www.fhwa.dot.gov/bridge/mtguide.pdf> – federal highway administration, deste arquivo foram raspadas diferentes tabelas para alimentação de nossos dados principais. Aqui foram raspadas diferentes tabelas.

**Coleta, Modelagem e Carga**

Uma vez definido o conjunto de dados, devemos coletar e armazená-los na nuvem, este processo de armazenagem segue as disposições de uma arquitetura para ETL, desenvolvendo assim está no Azure, utilizando a carga dos dados para o Data Warehouse/Data Lake. Utilizamos pipelines de ETL (Extração, Transformação e Carga) na Azure e Databricks. Criada conta de armazenamento com três camadas. Criado o pipeline. E criado nosso cluster com nosso notebook Na camada bronze foi colocado nossos dados brutos E posteriormente com o código chegamos até nossa camada silver com dados já previamente tratados A camada gold foi mais o cálculo e tratamento final dos dados para avaliação de risco em estruturas e disponibilizados para nossos clientes.

**Análise**

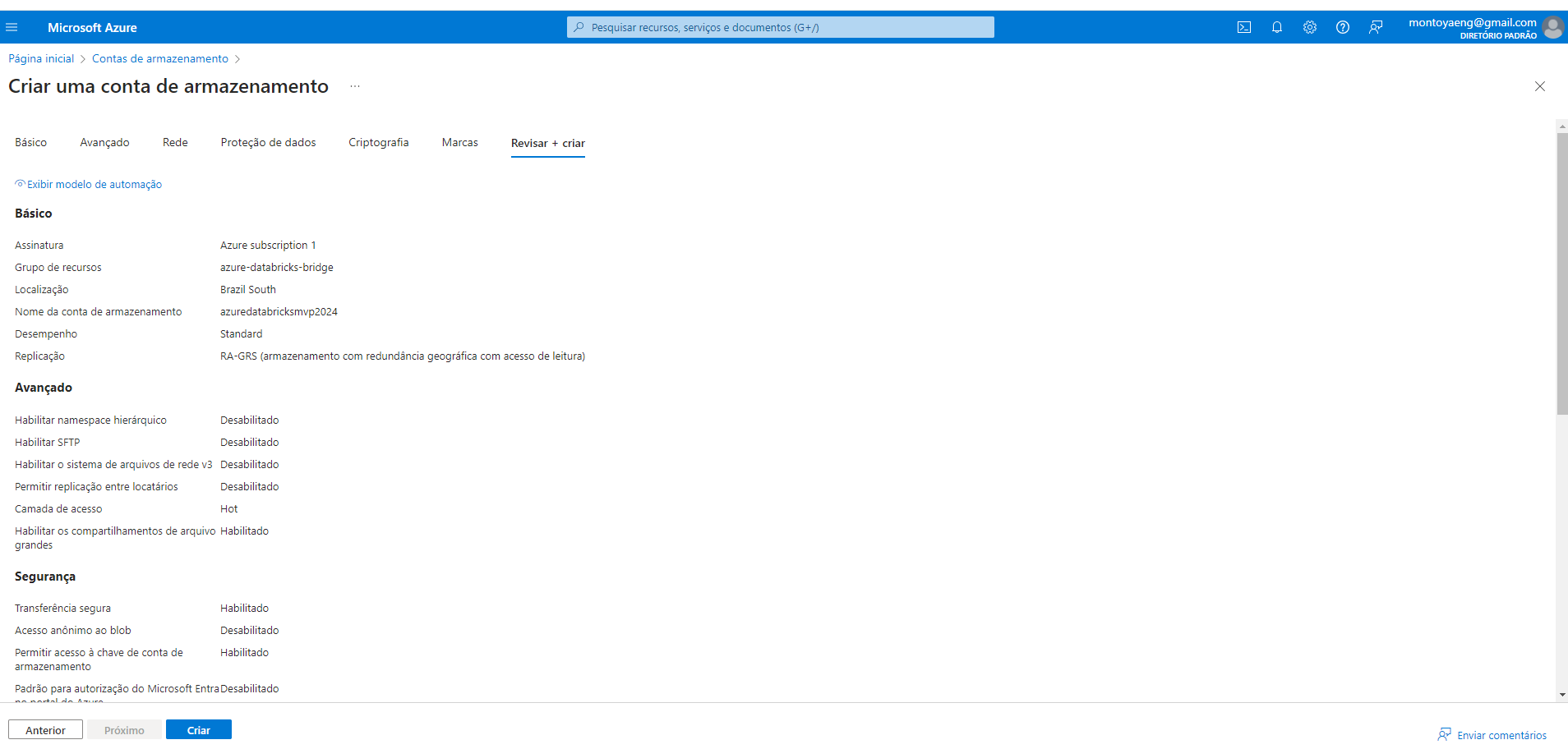
Qualidade de dados Os atributos encontrados tiveram alguns dados desnecessários para nossa análise, não é uma boa prática alterar a camada bruta, por isso que o tratamento dos dados é feito na silver, aqui deletamos dados que não seriam úteis para nossos questionamentos. Nas oportunidades de trabalho foi mais complexo o tratamento porque existem muitas funções e precisamos de avaliar as que são úteis para nosso questionamento.



2. Coleta, Modelagem e Carga

Uma vez definido o conjunto de dados, devemos coletar e armazená-los na nuvem, este processo de armazenagem segue as disposições de uma arquitetura par ETL, desenvolvendo assim está no Azure, utilizando a carga dos dados para o Data Warehouse/Data Lake. Utilizamos pipelines de ETL (Extração, Transformação e Carga) na Azure e Databricks.

Criada conta de armazenamento com as três camadas.



[

  {

    "id": "/subscriptions/2fd0ff2e-3433-4530-9e94-ed6191c23e33/resourceGroups/azure-databricks-bridge/providers/Microsoft.Resources/deployments/azuredatabricksmvp2024\_1719919771847/operations/4E4006BA29EEDDB8",

    "operationId": "4E4006BA29EEDDB8",

    "properties": {

      "provisioningOperation": "Create",

      "provisioningState": "Running",

      "timestamp": "2024-07-02T11:31:07.0479451Z",

      "duration": "PT5.0106552S",

      "trackingId": "53efd8f8-f03e-4fe2-a7e8-9496e97481d4",

      "statusCode": "Accepted",

      "targetResource": {

        "id": "/subscriptions/2fd0ff2e-3433-4530-9e94-ed6191c23e33/resourceGroups/azure-databricks-bridge/providers/Microsoft.Storage/storageAccounts/azuredatabricksmvp2024",

        "resourceType": "Microsoft.Storage/storageAccounts",

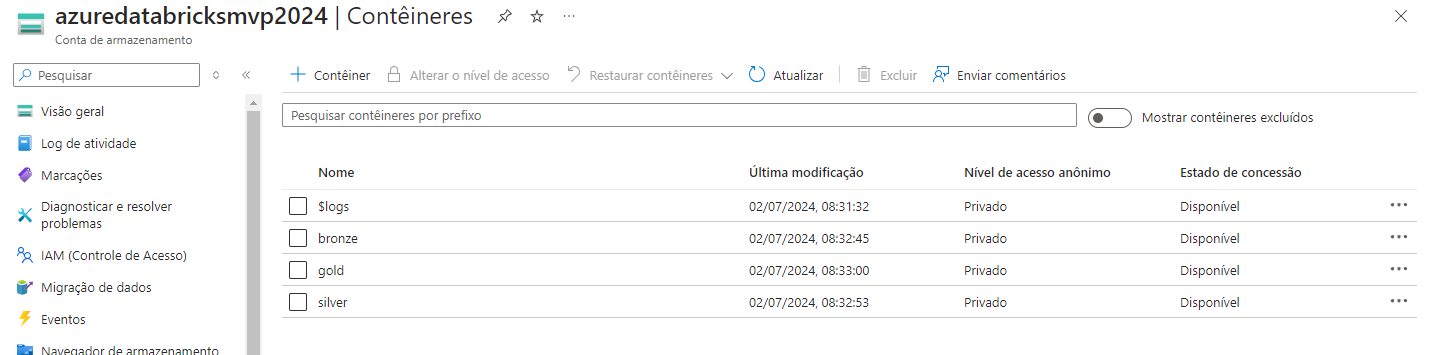
        "resourceName": "azuredatabricksmvp2024"

      }

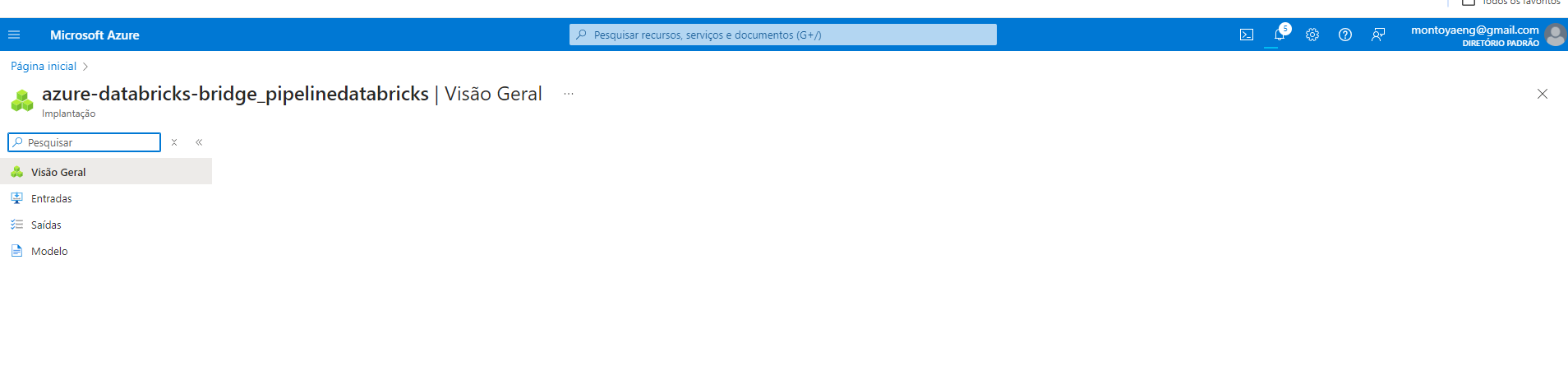
    }

  }

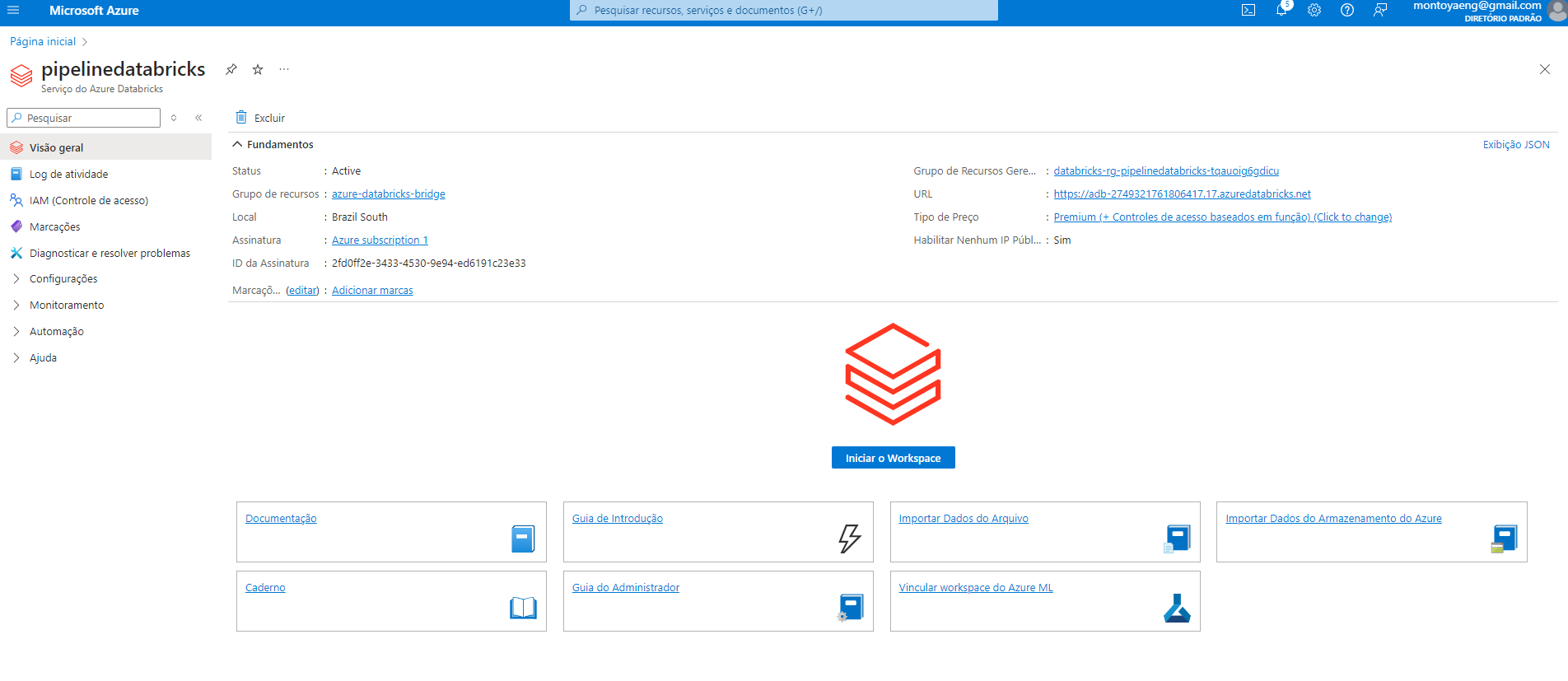
]



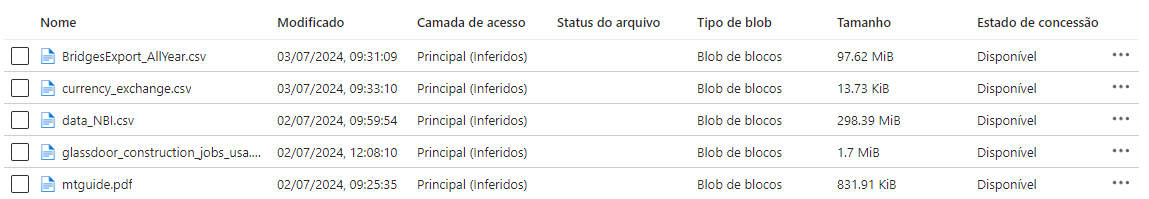
Criado o pipeline.



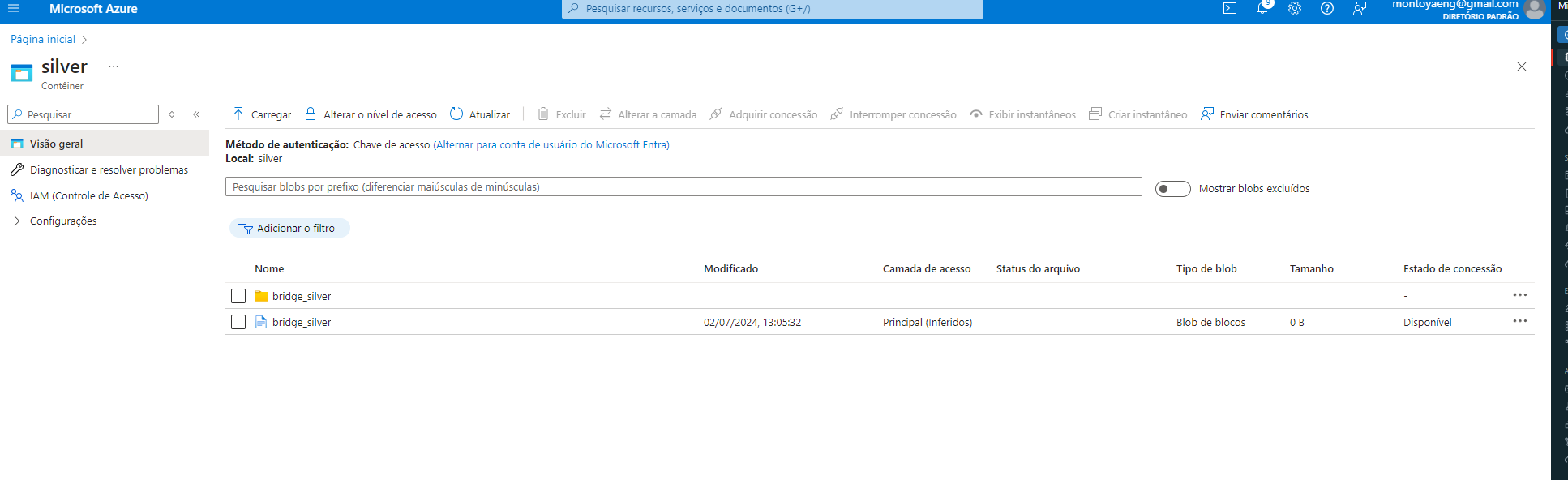
E criado nosso cluster com nosso notebook



Na camada bronze foi colocado nossos dados brutos



E posteriormente com o código chegamos até nossa camada silver com dados já previamente tratados



A camada gold foi mais o cálculo e tratamento final dos dados para avaliação de risco em estruturas e disponibilizados para nossos clientes

5. Análise

Qualidade de dados

Os atributos encontrados tiveram alguns dados desnecessários para nossa análise, não é uma boa pratica alterar a camada bruta, por isso que o tratamento dos dados é feito na silver, aqui deletamos dados que não seriam uteis para nossos questionamentos.

Nas oportunidades de trabalho foi mais complexo o tratamento porque existem muitas funções e precisamos de avaliar as que são uteis para nosso questionamento.

Solução do problema no arquivo databricks

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Interface gráfica do usuário, Tabela

Descrição gerada automaticamente

Uma imagem contendo Texto

Descrição gerada automaticamente

Texto

Descrição gerada automaticamente

Tabela

Descrição gerada automaticamente

Tabela

Descrição gerada automaticamente

Texto

Descrição gerada automaticamente com confiança média

Interface gráfica do usuário

Descrição gerada automaticamente

Texto

Descrição gerada automaticamente

Tabela

Descrição gerada automaticamente

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Tabela

Descrição gerada automaticamente

Uma imagem contendo Texto

Descrição gerada automaticamente

Tabela

Descrição gerada automaticamente

Interface gráfica do usuário, Aplicativo, Tabela, Excel

Descrição gerada automaticamente

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

Tabela

Descrição gerada automaticamente com confiança média

Gráfico, Linha do tempo

Descrição gerada automaticamente

Gráfico, Gráfico de explosão solar

Descrição gerada automaticamente

Tela de computador com texto preto sobre fundo branco

Descrição gerada automaticamente

Interface gráfica do usuário

Descrição gerada automaticamente

Interface gráfica do usuário, Gráfico

Descrição gerada automaticamente

Gráfico, Histograma

Descrição gerada automaticamente

Gráfico, Histograma

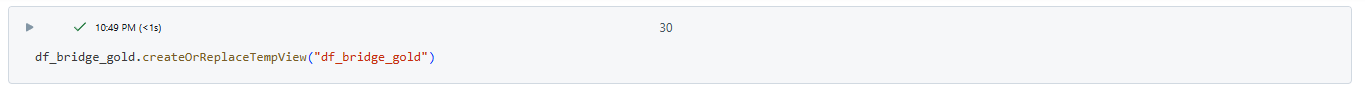
Descrição gerada automaticamente

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

Uma imagem contendo Tabela

Descrição gerada automaticamente



Texto

Descrição gerada automaticamente

Gráfico, Gráfico de barras

Descrição gerada automaticamente

Texto

Descrição gerada automaticamente