PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Gabriel da Silva Gonçalves

Construção de um pipeline de dados envolvendo busca, coleta, modelagem, carga e análise dos dados.
Análise sobre tiroteios policiais fatais nos Estados Unidos desde 2015, da base de dados “Fatal Encounters”

PROJETO DE PÓS-GRADUAÇÃO EM CIÊNCIA DE DADOS E ANALYTICS
APRESENTADO AO DEPARTAMENTO RESPONSÁVEL
DA PUC-RIO, COMO PARTE DOS REQUISITOS PARA OBTENÇÃO
DO DIPLOMA DE PÓS-GRADUAÇÃO LATO SENSU

Rio de Janeiro, 01 de Outubro de 2023.

1. Busca pelos dados

A busca pela base de dados foi feita no Kaggle.

O link é: Fatal US Police Violence (kaggle.com)

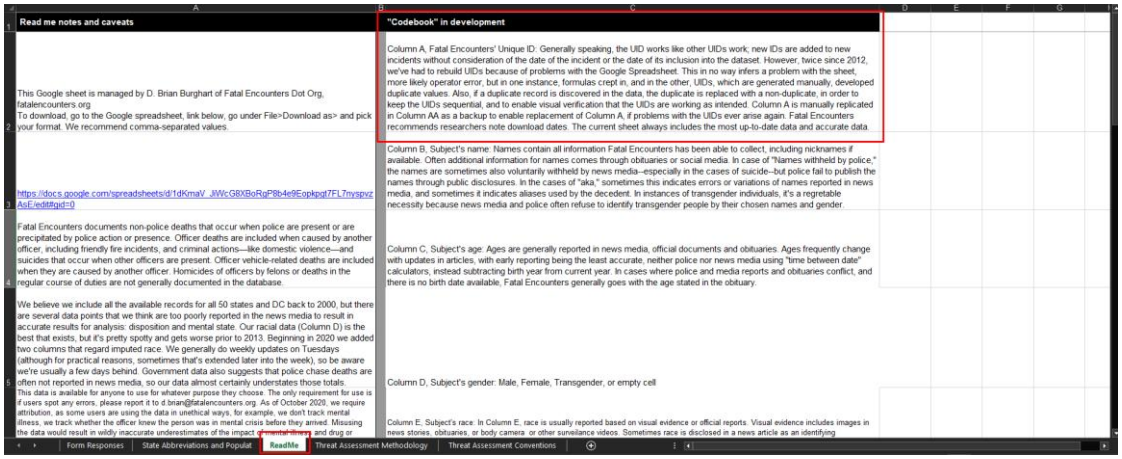
Na verdade, por se tratar de uma base bastante simples e que já tratada, foi optado por extrair de onde esses dados vieram, ou seja, do site: Fatal Encounters – A step toward creating an impartial, comprehensive and searchable national database of people killed during interactions with police.

O download foi feito pelo botão “Download FE Database”

Descrição da base:

“Fatal Encounters é um banco de dados de incidentes em que um indivíduo morre durante um encontro com policiais. A maioria dos encontros ocorre como resultado de homicídio policial, como quando os policiais atiram em uma pessoa que representa uma ameaça letal para eles ou outras pessoas. No entanto, existem outros tipos de encontros que não envolvem atos de homicídio policial, mas nos quais a polícia está envolvida ou presente. Os exemplos incluem um acidente de veículo durante uma perseguição policial ou um suicídio em uma situação de barricada.”
(informação retirada da planilha “FATAL ENCOUNTERS DOT ORG SPREADSHEET (See Read me tab).xlsx”)

A própria planilha oferece um dicionário de dados na aba “Read Me”, coluna “Codebook in development”.



As suas abas mais relevantes são: “Form Responses”, onde os dados de fato estão e “State Abbreviations and Populat”, em que há dados sobre a população de cada estado.

2. Tratamento prévio

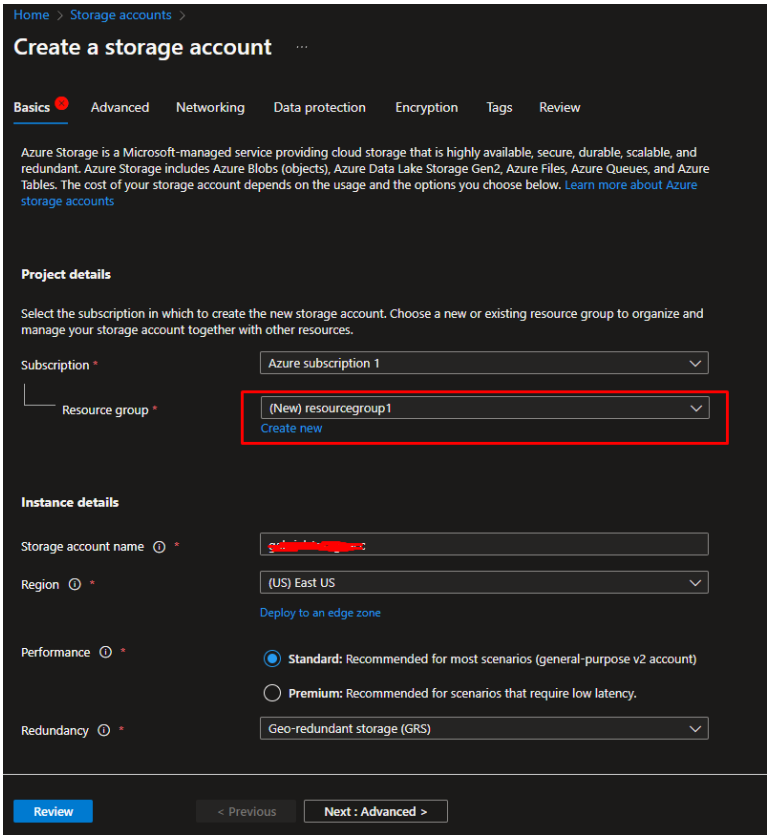
A fim de criar arquivos csv separados, cada um com o nome de uma aba, foi criado um script para fazer isso. Esse script, no Github, é chamado de “separar_abas.py”.

Foi feito um tratamento prévio nos dados a fim de tornar o seu upload mais fácil para a nuvem. Esse processo está melhor descrito no arquivo “analise_inicial.ipynb”

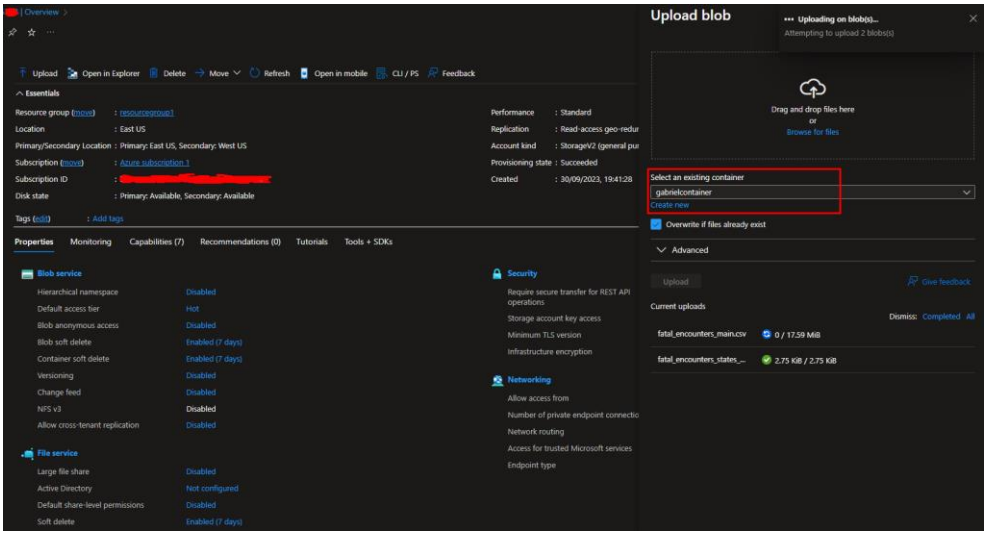
3. Uso da nuvem (Azure)

Por já possuir familiaridade com a plataforma, foi optado por se utilizar o Microsoft Azure com o intuito de disponibilizar a base de dados em um banco SQL e usar o Databricks posteriormente.

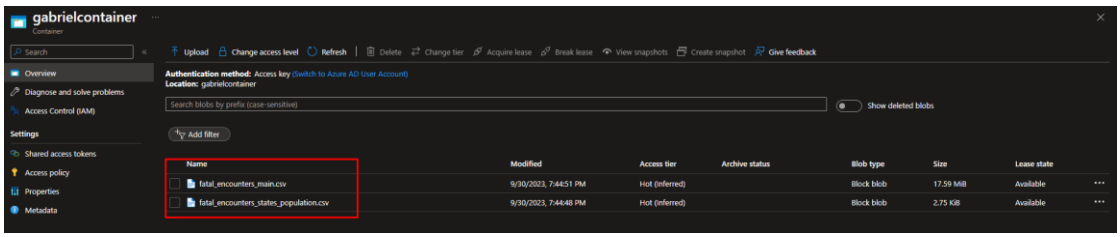
Após criar uma conta free trial no Azure, foi, primeiramente, necessário criar um storage account e um resource group:



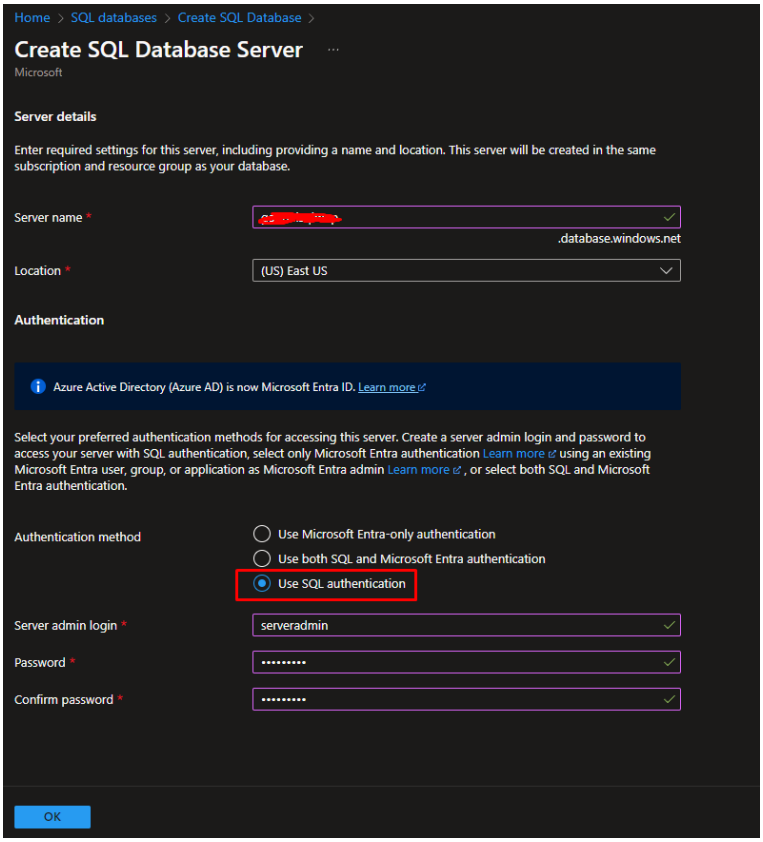
Em seguida, dentro desse storage account, foi feito o upload dos arquivos em csv, sendo também necessário criar um container:

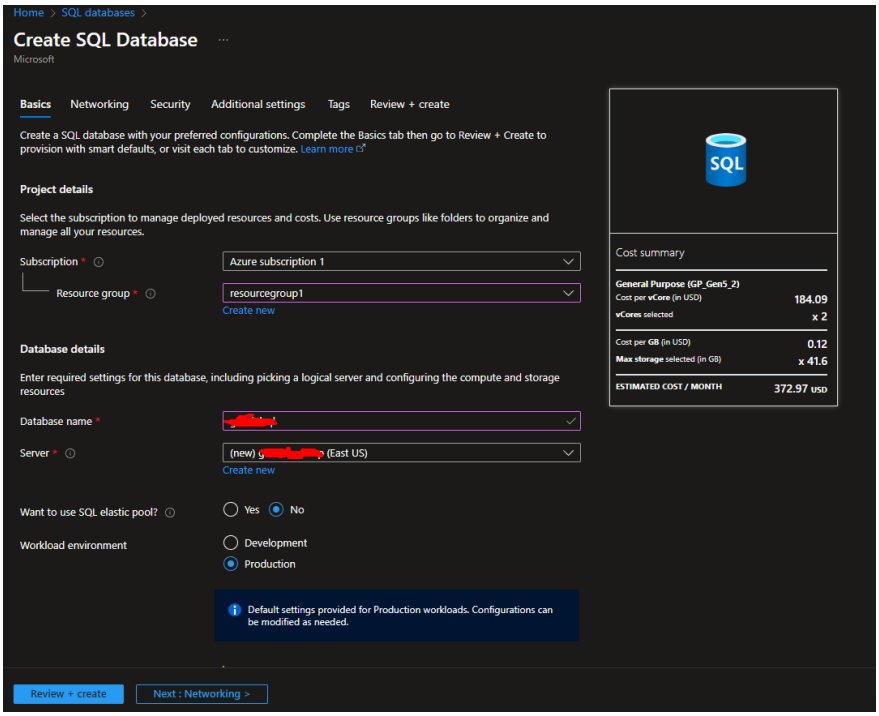


Dessa forma, os arquivos ficaram disponíveis:

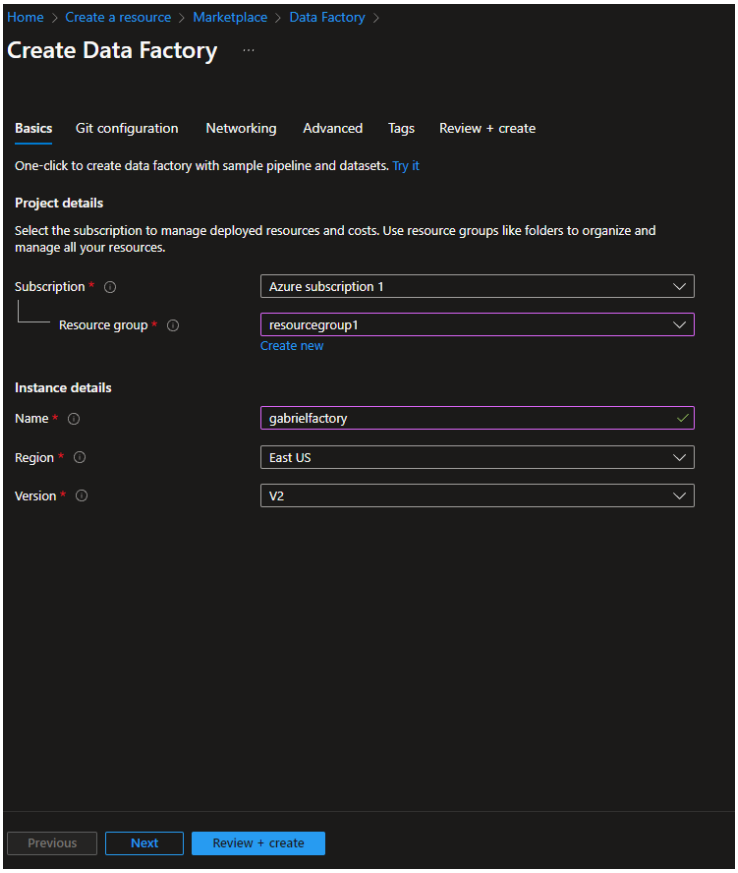


Tendo os arquivos, partiu-se para a etapa de criação do servidor do banco de dados SQL:

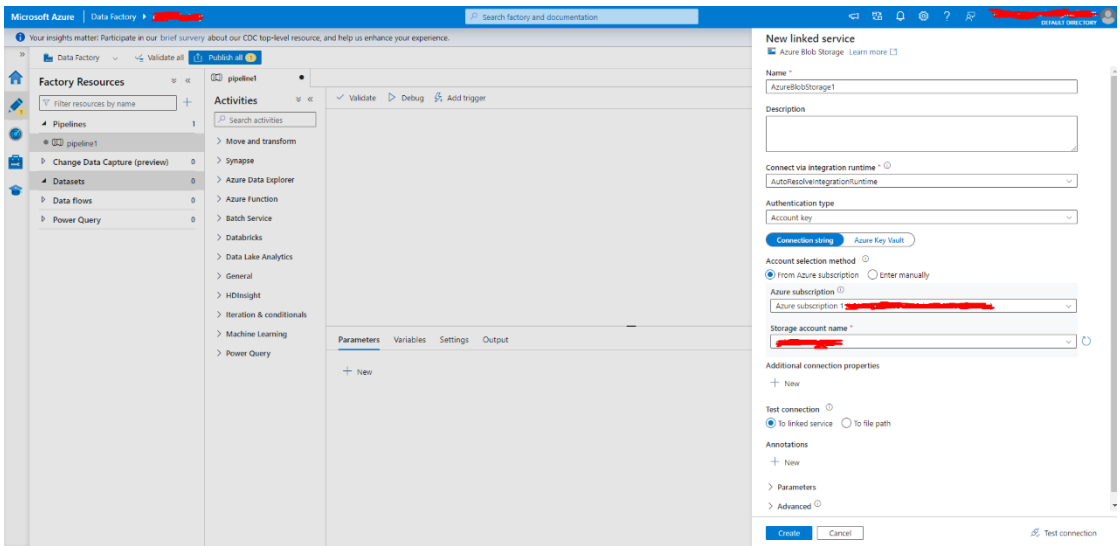




Com isso, foi criado um Data Factory para, posteriormente, ser realizado o processo de pipeline:



Ao acessar o Data Factory, a etapa de criação do pipeline foi realizada e envolveu: ingestão das tabelas presentes no blob storage, criação do dataflow e do pipeline.



Set properties

Name

DelimitedText1

Linked service *

AzureBlobStorage1

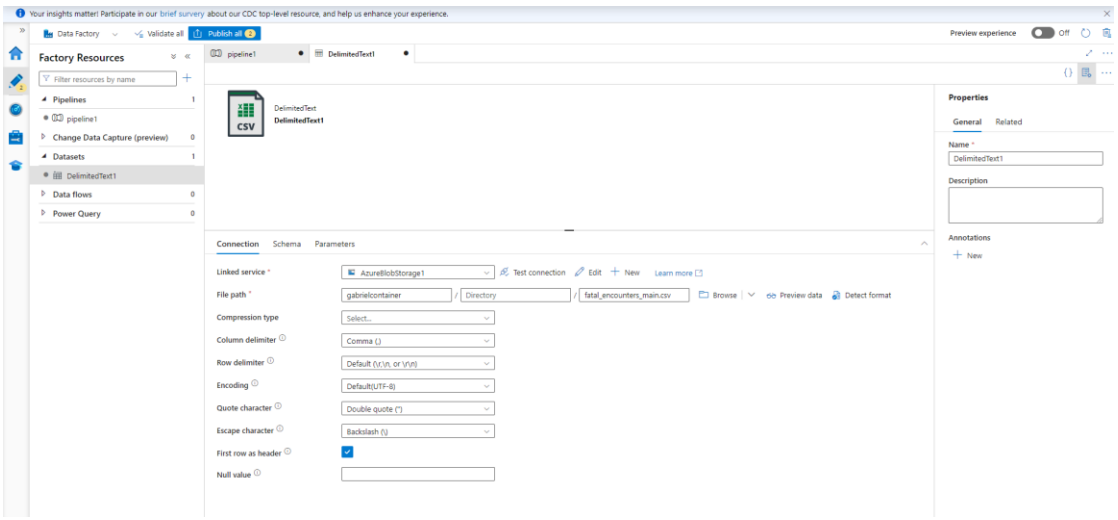
File path

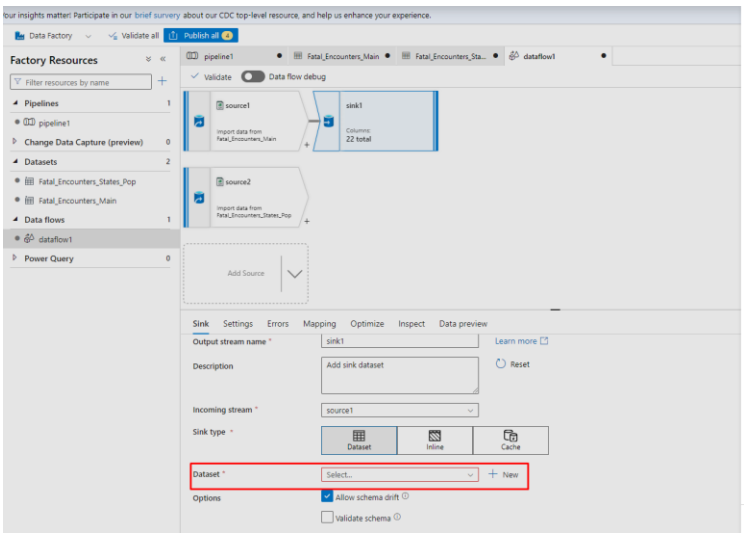
gabrielcontainer / Directory / fatal_encounters_main....

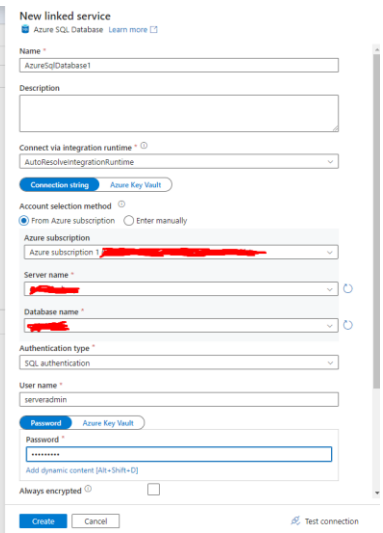
First row as header

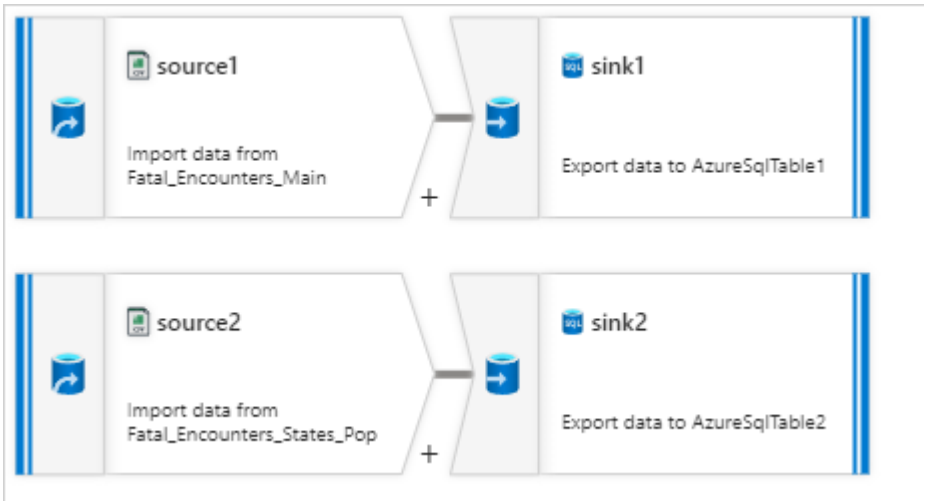
Import schema

From connection/store From sample file None

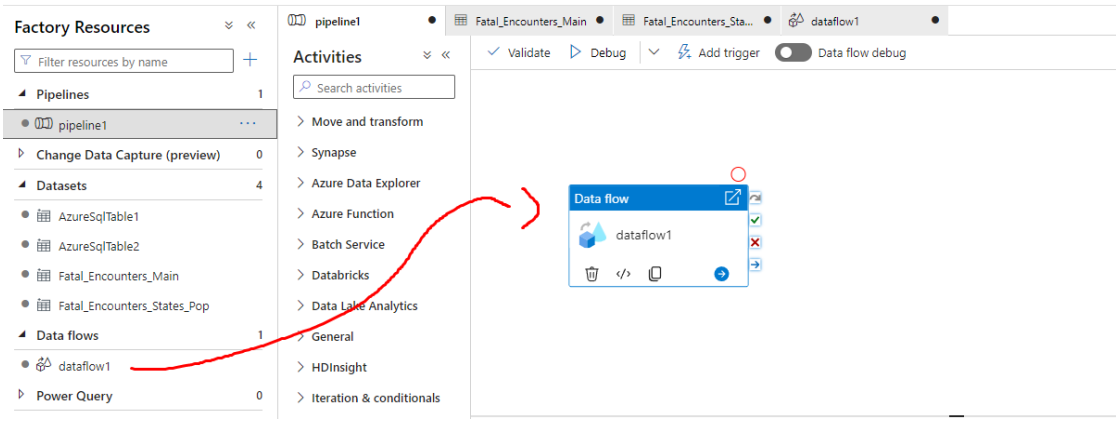




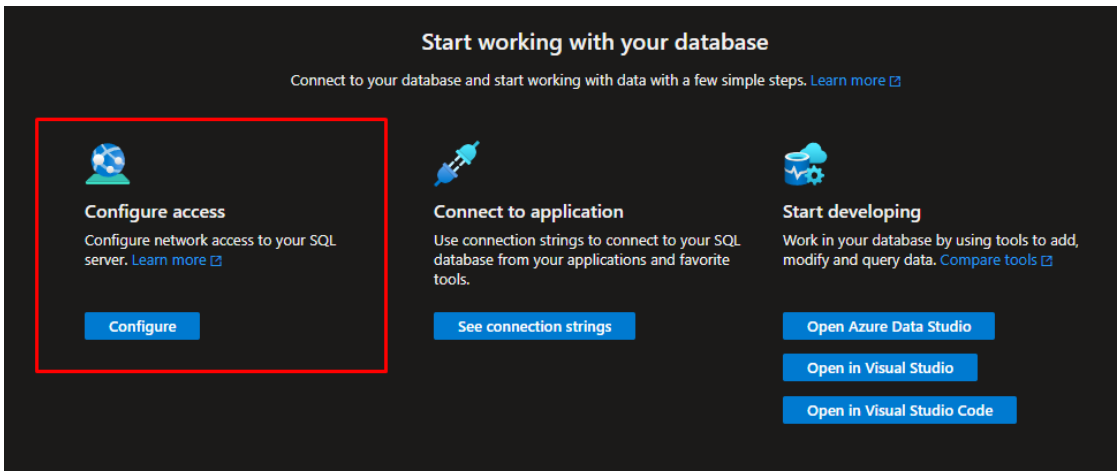




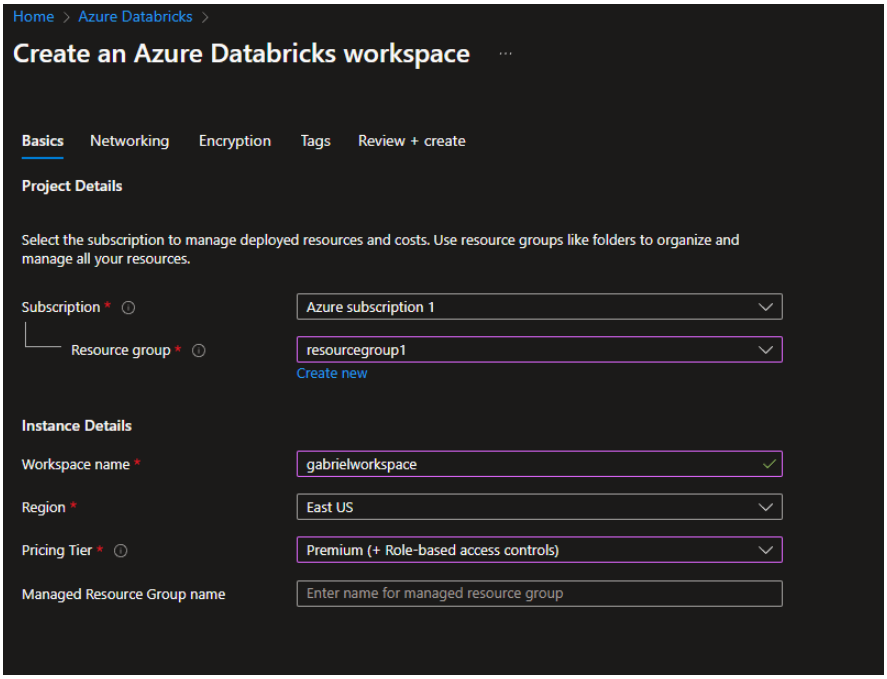
A imagem acima representa o dataflow, ou seja, não foram feitas alterações nas tabelas. As duas foram, individualmente, exportadas para o servidor do SQL por meio do “sink”. E esse dataflow foi usado para criar o pipeline:



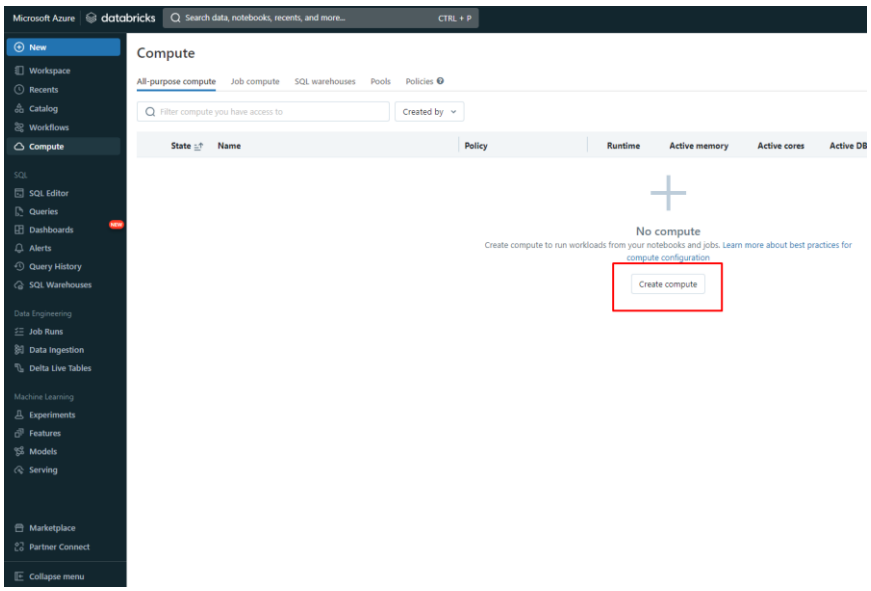
Contudo, para que o processo ocorresse sem erros foi antes necessário configurar o acesso ao servidor SQL:

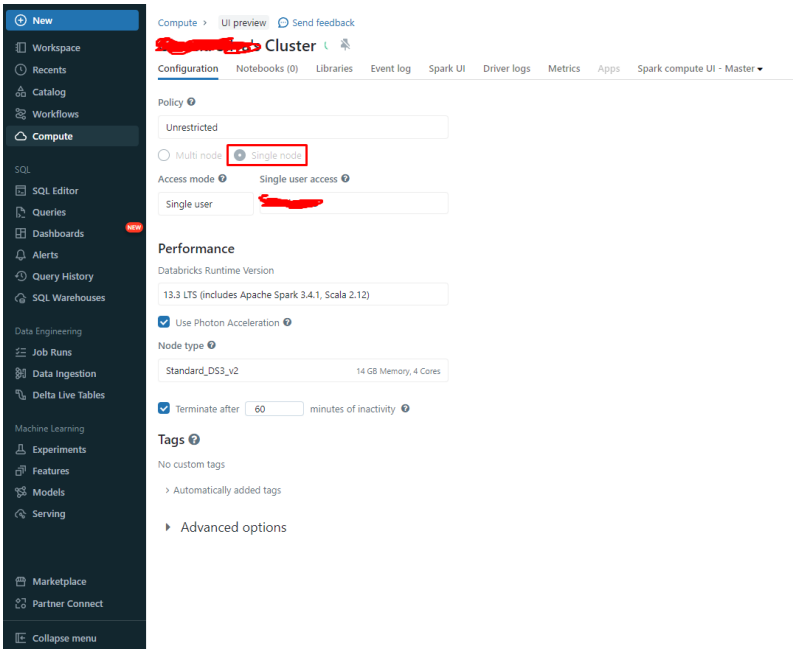


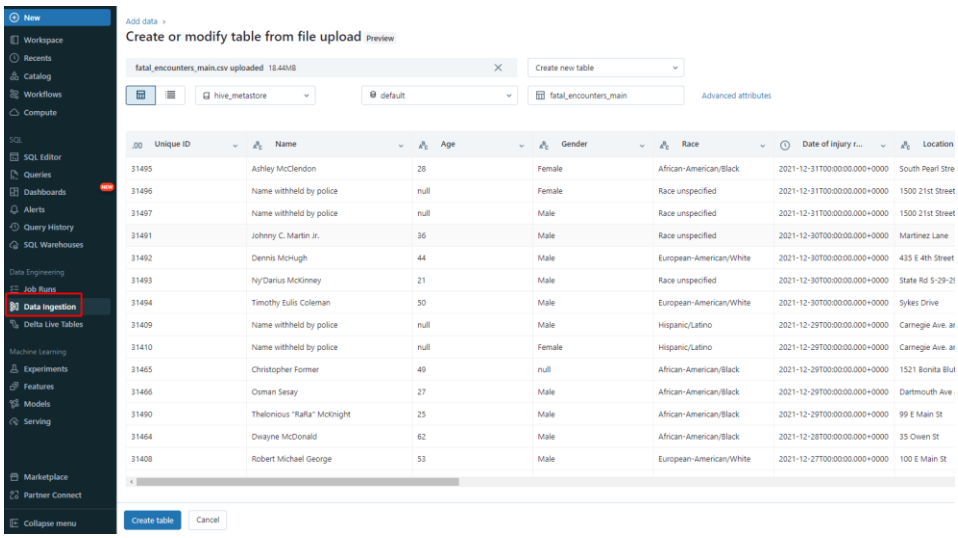
Após a criação e configuração do servidor SQL, o pipeline pôde ser rodado. Logo em seguida, foi criado um ambiente Databricks:



Foi criado um cluster para poder ser usado no notebook para as análises:







4. Power BI

Por meio da base de dados do SQL Server, foi gerado um report interativo no Power BI com os dados de “Fatal Encounters”.

Link para o relatório

5. Análise

A parte de análise, de maneira geral, está presente no arquivo “Análise de Fatal Encounters.ipynb”, mas nesta seção serão detalhados pontos adicionais.

a. Qualidade dos dados

Algumas colunas possuíam dados inconsistentes, mas como as análises envolviam, individualmente, uma ou poucas colunas, isso foi resolvido e tratado em cada análise como pode ser visto no arquivo de análise.

b. Solução do problema

Os problemas, de maneira geral, foram resolvidos e podem ser vistos no arquivo de análise.