

MVP - SPRINT 03 - Engenharia de Dados

Aluno: Daniel de Almeida Azevedo Silva

Professores: Sérgio Lifschitz, Profa. Dra. Fernanda Baião, Prof. Msc. Antony Seabra, Marcos Villas , Victor Teixeira de Almeida e Silvio.

Relatório da Construção do Pipeline no Azure Databricks da Microsoft para Análise dos Índices da Segurança Pública do Estado do Rio de Janeiro

1 – Objetivo

Analisar índices de segurança pública do estado do Rio de Janeiro disponíveis no site <https://www.ispdados.rj.gov.br/> , consultados em 01/07/2024, e responder às perguntas abaixo:

- A) O programa de implantação de UPPs em algumas comunidades do estado do Rio de Janeiro foi preponderante para a diminuição da violência no estado do Rio de Janeiro?
- B) Os índices de violência diminuíram na comunidade Santa Marta após a implantação da upp em dezembro de 2008?
- C) Vila Kennedy é a região mais violenta do Rio de Janeiro, apesar da upp instalada?

2 - Coleta dos Dados

Utilizei duas tabelas com informações sobre índices de criminalidade nas regiões das UPPs, disponibilizadas pelo ISP.

Para extrair os dados destas para o datawarehouse no databrick, segui os passos abaixo:






2.1 – Criada conta FREE TRIAL para acesso às ferramentas nuvem AZURE da Microsoft. Duração: 15 dias

2.2 – Criação da Databricks workspace

Microsoft Azure

Upgrade

Search resources, services, and docs (G+)



danielalmeidaas@gmail...
DIRETÓRIO PADRÃO

[Home](#) > [All resources](#) > [Create a resource](#) > [Marketplace](#) > [Azure Databricks](#) >

Create an Azure Databricks workspace

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription *

Free Trial

Resource group *

(New) rgmvpsprint3ebdfree

[Create new](#)

Instance Details

Workspace name *


wsmvpazdatabrickpremiumbra


Region *

Brazil South

Pricing Tier *

Premium (+ Role-based access controls)

 We selected the recommended pricing tier for your workspace. You can change the tier based on your needs.



Managed Resource Group name

mrgmvp

[Review + create](#)

[< Previous](#)

[Next : Networking >](#)

Problemas ao tentar criar com Pricing tier também Free. Funcionou com a Pricing Tier sugerida pela Microsoft: Premium (+ role-base access controls).

[Home](#) > [All resources](#) > [Create a resource](#) > [Marketplace](#) > [Azure Databricks](#) >

Create an Azure Databricks workspace ...

Validation Succeeded

[Basics](#) [Networking](#) [Encryption](#) [Security & compliance](#) [Tags](#) [Review + create](#)

Summary

Basics

Workspace name	wsmvpazdatabrickpremiumbra
Subscription	Free Trial
Resource group	rgmvpsprint3ebdfree
Region	Brazil South
Pricing Tier	premium
Managed Resource Group name	mrgmvp

Networking

Deploy Azure Databricks workspace with Secure Cluster Connectivity (No Public IP) ☐ No

[Create](#)[< Previous](#)[Download a template for automation](#)

2.3 – Criando Cluster (poder de processamento)

Compute >

Daniel Almeida's Cluster

More

Terminate

Configuration

Notebooks (0)

Libraries

Event log

Spark UI

Driver logs

Metrics

Apps

Spark compute UI - Master

Policy

Unrestricted

Multi node

Single node

Access mode

No isolation shared

Performance

Databricks Runtime Version

15.3 (includes Apache Spark 3.5.0, Scala 2.12)

Use Photon Acceleration

Worker type

Min workers

Max workers

Current

Standard_D4ads_v5

16 GB Memory, 4 Cores

2

8

0

Spot instances

Driver type

Standard_DS3_v2

14 GB Memory, 4 Cores

Summary

2-8 Workers

32-128 GB Memory

8-32 Cores

1 Driver

14 GB Memory, 4 Cores

Runtime

15.3.x-scala2.12

Photon

Standard_D4ads_v5

Standard_DS3_v2

5-18 DBU/h

2.4 – Requisição de aumento do número de virtual cpus

Microsoft Azure

Search resources, services, and docs (G+)

Home >

...

New Quota Request

Refresh

Download

You can now set up alerts for your Quota usage and receive notifications. Simply click on any Quota to create one. [Learn More.](#)

Recommended

To view and manage quotas across all your subscriptions from a central location, go to [Azure Quotas.](#)

Search

Provider : Compute

Region : All

Usage : Show all

Showing 1 to 100 of 6892 records in 1 groups.

Quota name	Region	Subscription
No usage (Showing 100 of 6892)		
<input type="checkbox"/> Availability Sets	Australia Central	DEVELOPER - Azure subscri
<input type="checkbox"/> Total Regional vCPUs	Australia Central	DEVELOPER - Azure subscri
<input type="checkbox"/> Virtual Machines	Australia Central	DEVELOPER - Azure subscri
<input type="checkbox"/> Virtual Machine Scale Sets	Australia Central	DEVELOPER - Azure subscri
<input type="checkbox"/> Dedicated vCPUs	Australia Central	DEVELOPER - Azure subscri

< Previous

Page 1 of 69

Next >

New Quota Request

Successful 1

Partial increase 0

Unsuccessful 0

We have adjusted your quota.

DEVELOPER - Azure subscription 1 - 97 USD mes

Received (vCPU)

New limit (vCPU)

Brazil South

Standard DDSv5 Family vCPUs

8 of 8

8

Microsoft Azure

Search resources, services, and docs (G+)

Home >

...

New Quota Request

Refresh

Download

You can now set up alerts for your Quota usage and receive notifications. Simply click on any Quota to create one. [Learn More.](#)

Recommended

To view and manage quotas across all your subscriptions from a central location, go to [Azure Quotas.](#)

Search

Provider : Compute

Region : All

Usage : Show all

Showing 1 to 100 of 6892 records in 4 groups.

Quota name	Region	Subscription
Usage at or near quota (1)		
<input type="checkbox"/> Standard DDSv5 Family vCPUs	Brazil South	DEVELOPER - Azure subscri
Usage at regular level (1)		
<input type="checkbox"/> Total Regional vCPUs	Brazil South	DEVELOPER - Azure subscri
Usage at low level (2)		

< Previous

Page 1 of 69

Next >

New Quota Request

Successful 1

Partial increase 0

Unsuccessful 0

We have adjusted your quota.

DEVELOPER - Azure subscription 1 - 97 USD mes

Received (vCPU)

New limit (vCPU)

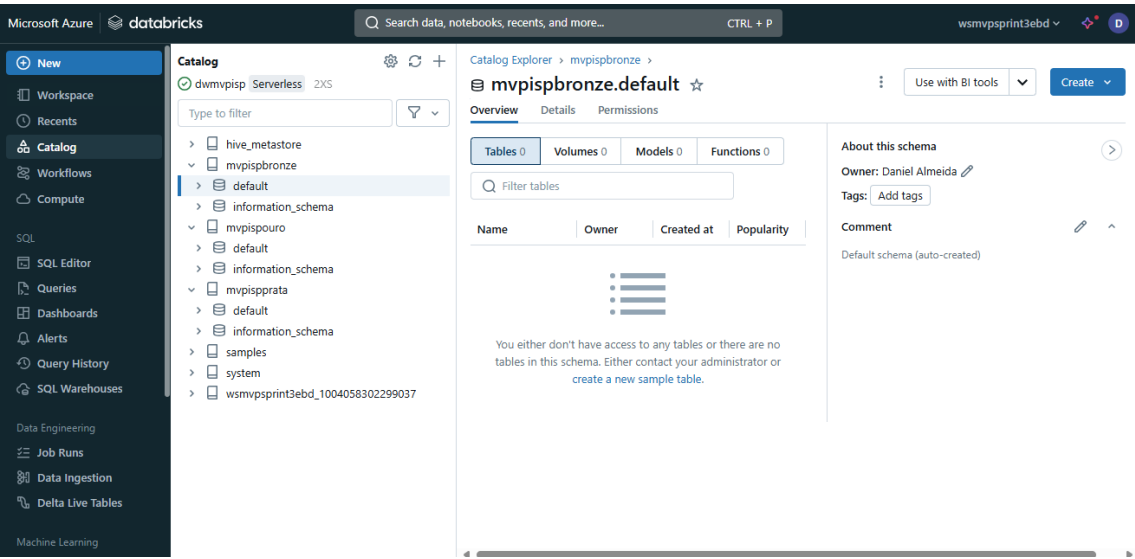
Brazil South

Standard DDSv5 Family vCPUs

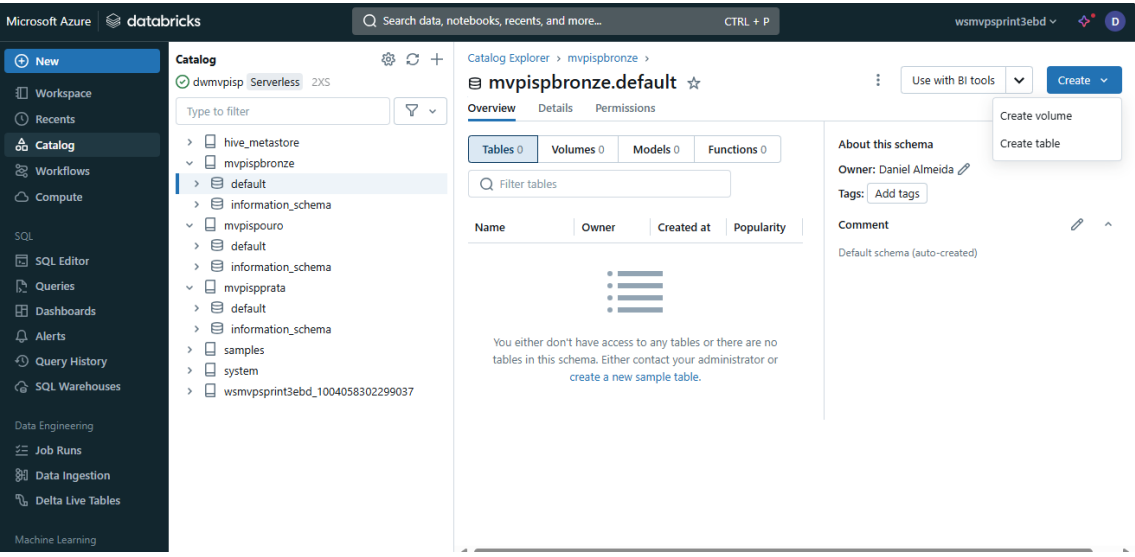
8 of 8

16

2.5 – Criei os bancos de dados (catalogs) mvvispbronze, mvvispprata e mvvispouro no databricks, visando respeitar respectivamente as camadas bronze, prata e ouro.



2.6 – Criação da Tabela de UPPs a partir do upload do arquivo CSV no databricks



Microsoft Azure databricks

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

wsmvpsprint3ebd 2XS

New

Workspace

Recents

Catalog

Workflows

Compute

SQL

SQL Editor

Queries

Dashboards

Alerts

Query History

SQL Warehouses

Data Engineering

Job Runs

Data Ingestion

Delta Live Tables

Machine Learning

Add data >

Create or modify table from file upload

dwmpisp Serverless 2XS

DATA

Drop one or more files here, or [browse](#)

Maximum of 10 files and total upload size of 2GB

Requires a SQL warehouse or a cluster with Databricks Runtime 10.3 and above

Supported file formats: .csv, .tsv, .tab, .json, .jsonl, .avro, .parquet, .txt, or .xml

For larger files, for other file formats, or for uploading files to a non-tabular dataset without creating a table, [upload to a Volume in Unity Catalog](#).

Microsoft Azure databricks

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

wsmvpsprint3ebd 2XS

New

Workspace

Recents

Catalog

Workflows

Compute

SQL

SQL Editor

Queries

Dashboards

Alerts

Query History

SQL Warehouses

Data Engineering

Job Runs

Data Ingestion

Delta Live Tables

Machine Learning

Add data >

Create or modify table from file upload

dwmpisp Serverless 2XS

AreasUPP_utf-8.csv

Preview mode

Advanced attributes

Automatically detect file type

File type

CSV

Column delimiter

;

Escape character

"

First row contains the header

Automatically detect column types

Rows span multiple lines

Merge the schema across multiple files

Cancel

Create table

Microsoft Azure databricks

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

wsmvpsprint3ebd 2XS

New

Workspace

Recents

Catalog

Workflows

Compute

SQL

SQL Editor

Queries

Dashboards

Alerts

Query History

SQL Warehouses

Data Engineering

Job Runs

Data Ingestion

Delta Live Tables

Machine Learning

Add data >

Create or modify table from file upload

dwmpisp Serverless 2XS

AreasUPP_utf-8.csv uploaded 2.52KB

Create new table

Preview mode

Catalog

Schema

Table name

areas_upp_utf_8

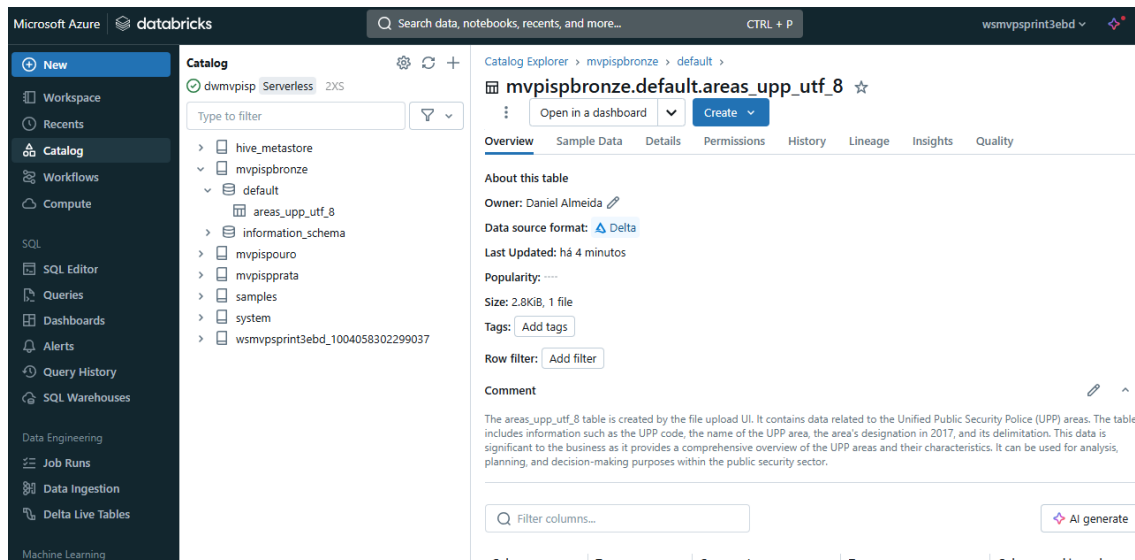
Advanced attributes

Previewing 38 rows, 4 columns

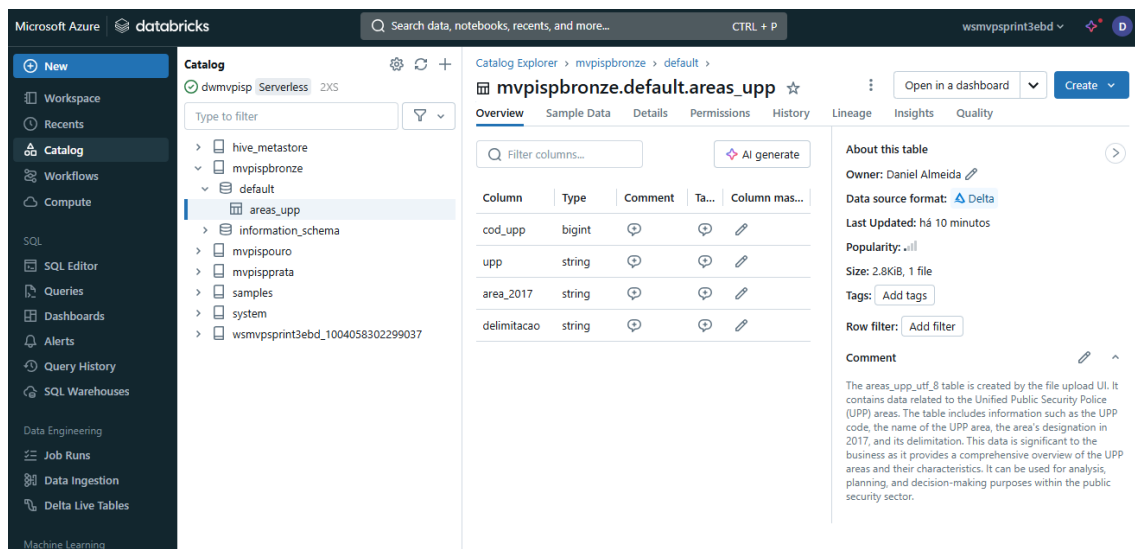
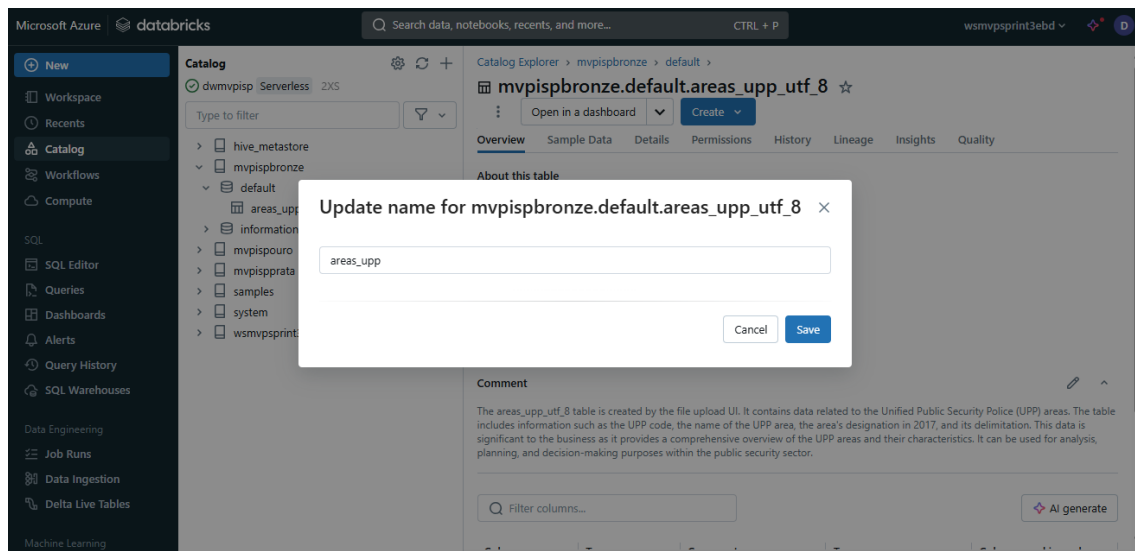
	cod_upp	upp	area_2017	delimitacao
1		Santa Marta	96.796,45	DOERJ nº 044 de 10/03/2011, DOERJ nº 1...
2		Cidade de Deus	2.394.810,59	DOERJ nº 039 de 28/02/2011
3		Batam	1.183.698,49	DOERJ nº 044 de 10/03/2011 e DOERJ nº ...
4		Chapéu Mangueira / Babilônia	172.161,69	DOERJ nº 044 de 10/03/2011
5		Pavão-Pavãozinho	248.637,83	DOERJ nº 044 de 10/03/2011
6		Tabajaras	284.557,86	DOERJ nº 044 de 10/03/2011
7		Providência	679.408,73	DOERJ nº 044 de 10/03/2011
8		DOERJ nº 044 de 10/03/2011

Cancel

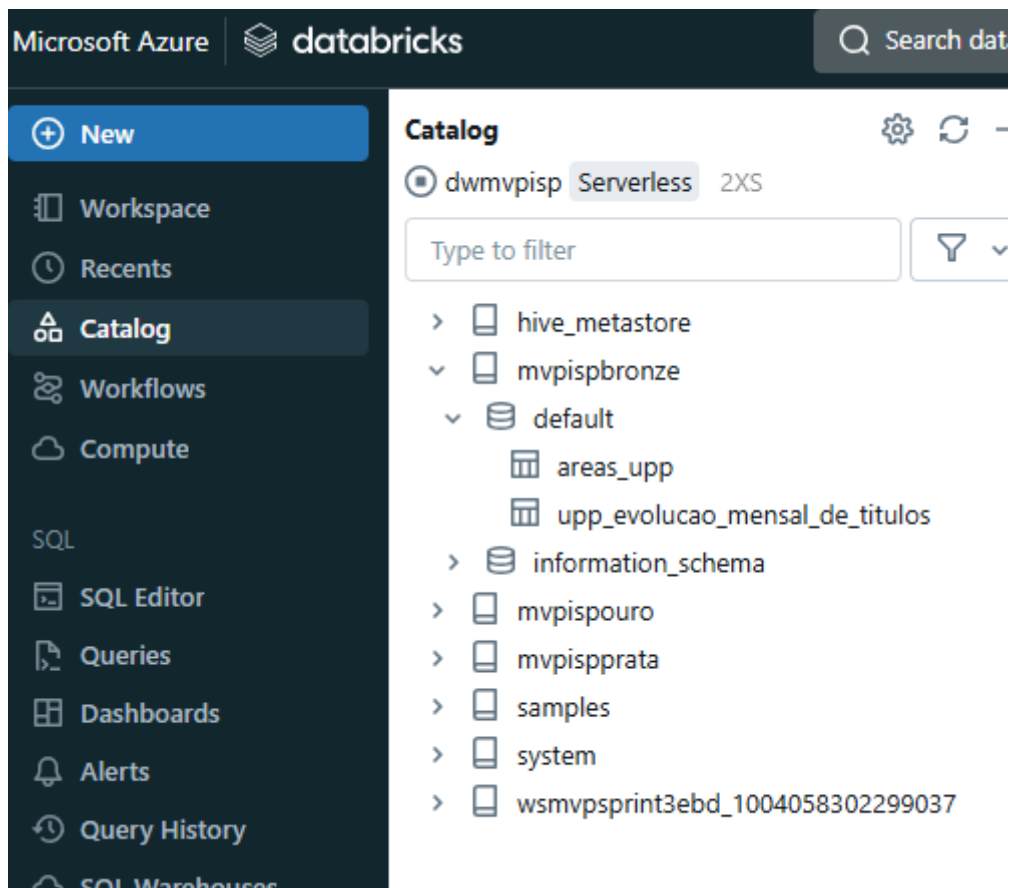
Create table



Renomeando a tabela para Areas_UPP, removendo o sufixo _utf_8.



A tabela com os índices criminais passou pelo mesmo processo.



3- Modelagem

Utilizei duas tabelas com informações sobre índices de criminalidade nas regiões das UPPs, disponibilizadas pelo ISP.

A primeira tabela `areas_upp` já se encontra normalizada. Segue abaixo o dicionário de dados:

Dicionário de Dados da Tabela AREAS_UPP

COLUNA	TIPO	Descrição
<code>cod_upp</code>	inteiro	Código da upp
<code>upp</code>	texto	Nome da upp
<code>area_2017</code>	numerico	Área da região da UPP

delimitacao	texto	Delimitação da UPP, coforme DOERJ
data_implantacao	Data	Data da Implantação da UPP
zona	texto	Zona em que a comunidade está localizada

Segue abaixo o dicionário de dados da segunda tabela, que contém os índices de criminalidade mensais das UPPs de 2007 a 2021:

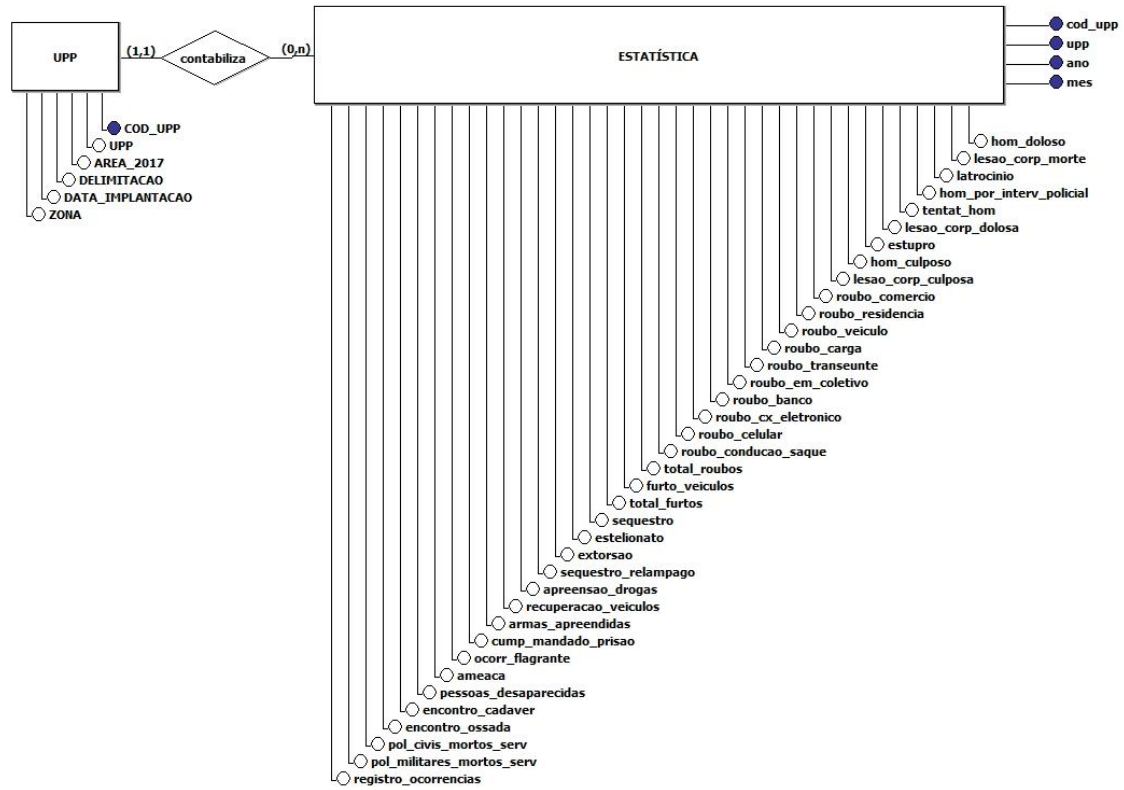
Dicionário de Dados da tabela upp_estatísticas

Coluna	Tipo	Descrição
cod_upp	inteiro	Código da UPP
ano	inteiro	Ano dos índices de criminalidade
mês	inteiro	Mês dos índices de criminalidade
hom_doloso	inteiro	Homicídio doloso
lesao_corp_morte	inteiro	Lesão corporal seguida de morte
latrocinio	inteiro	Latrocínio (roubo seguido de morte)
hom_por_interv_policial	inteiro	Morte por intervenção de agente do Estado
tentat_hom	inteiro	Tentativa de homicídio
lesao_corp_dolosa	inteiro	Lesão corporal dolosa
estupro	inteiro	Estupro
hom_culposo	inteiro	Homicídio culposo
lesao_corp_culposa	inteiro	Lesão corporal culposa
roubo_comercio	inteiro	Roubo a estabelecimento comercial
roubo_residencia	inteiro	Roubo a residência
roubo_veiculo	inteiro	Roubo de veículo
roubo_carga	inteiro	Roubo de carga
roubo_transeunte	inteiro	Roubo a transeunte
roubo_em_coletivo	inteiro	Roubo em coletivo
roubo_banco	inteiro	Roubo a banco
roubo_cx_eletronico	inteiro	Roubo de caixa eletrônico
roubo_celular	inteiro	Roubo de aparelho celular
roubo_conducao_saque	inteiro	Roubo com condução da vítima para saque em I.F.
total_roubos	inteiro	Roubos (soma o número de ocorrências de todos os tipos de roubo, inclusive outros roubos que não estão discriminados aqui)
furto_veiculos	inteiro	Furto de veículo
total_furtos	inteiro	Furtos (soma o número de ocorrências de todos os tipos de furto, inclusive outros furtos que não estão discriminados aqui)

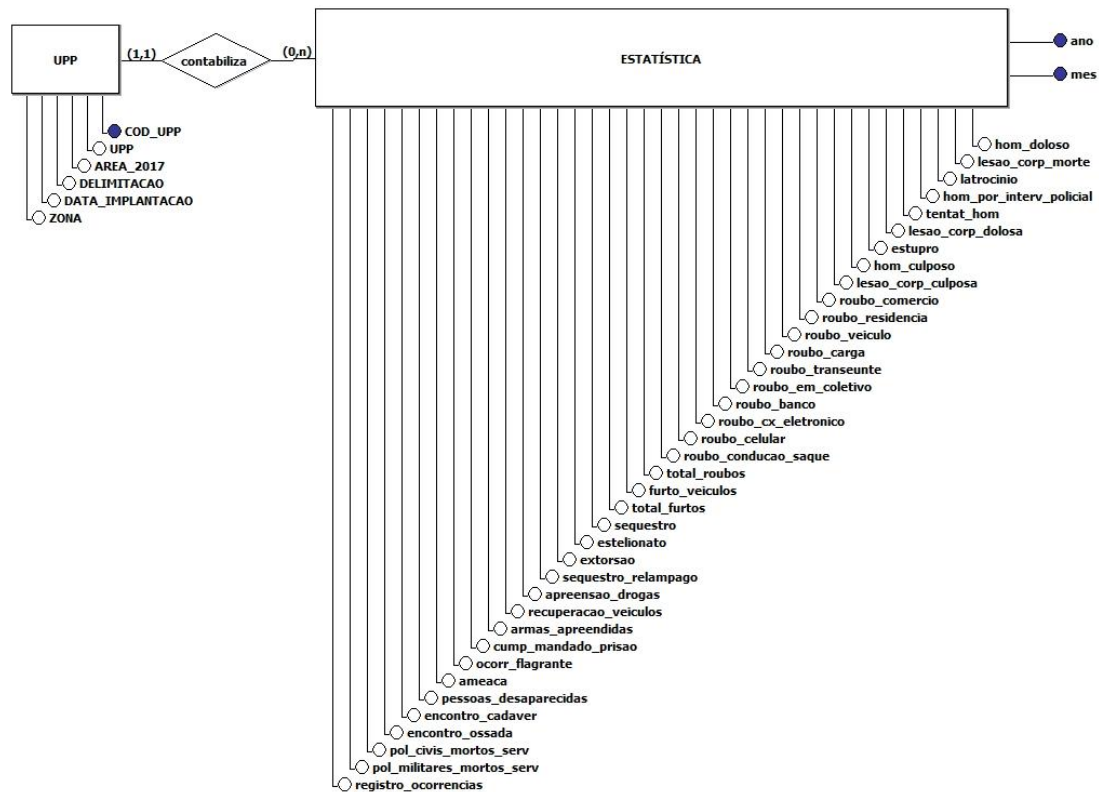
sequestro	inteiro	Extorsão mediante sequestro (sequestro clássico)
extorsao	inteiro	Extorsão
sequestro_relampago	inteiro	Extorsão com momentânea privação da liberdade (sequestro relâmpago)
estelionato	inteiro	Estelionato
apreensao_drogas	inteiro	Apreensão de drogas
recuperacao_veiculos	inteiro	Recuperação de veículo
armas_apreendidas	inteiro	Armas apreendidas
cump_mandado_prisao	inteiro	Cumprimento de mandado de prisão
ocorr_flagrante	inteiro	Ocorrências com flagrante
ameaca	inteiro	Ameaça (vítimas)
peessoas_desaparecidas	inteiro	Pessoas desaparecidas
encontro_cadaver	inteiro	Encontro de cadáver
encontro_ossada	inteiro	Encontro de ossada
registro_ocorrencias	inteiro	Registro de ocorrências (total lavrado no mês)

A única transformação realizada foi normalizar a tabela dos índices criminais:

MODELO CONCEITUAL DESNORMALIZADO - CAMADA BRONZE



MODELO CONCEITUAL NORMALIZADO - CAMADA OURO



4 – Análise dos Dados

4) A) Qualidade dos Dados

Visando o aumento da qualidade dos dados brutos carregados na camada bronze (banco de dados mvpsisbronze), na camada prata (banco de dados mvpsisprata) realizei algumas transformações para normalização do modelo relacional. Tal tratamento futuramente possibilitará ganho de performance.

A qualidade dos dados foi satisfatória, pois não foram detectados dados faltantes e nem outliers. No notebook databrick, cujo link foi disponibilizado no item seguinte, análise, disponibilizei um exemplo de teste da qualidade dos dados da coluna total_roubos.

4) B) Análise

Os questionamentos sobre a segurança pública do estado do Rio de Janeiro presentes no tópico objetivo deste trabalho foram respondidas através de consultas *sql* executadas na camada ouro (banco de dados mvpsisouro) e análise dos respectivos resultados.

A análise dos dados foi realizada na camada ouro. As evidências podem ser consultadas no notebook do databrick, exportado para html. Este pode ser acessado na url [ECDA_PUCRIO_232/MVP SPRINT03 ENGENHARIA DE DADOS Notebook Versão FINAL 2024-07-12.html at main · Dani31A1m3ida/ECDA_PUCRIO_232 \(github.com\)](https://github.com/Dani31A1m3ida/ECDA_PUCRIO_232/blob/main/ECDA_PUCRIO_232/MVP%20SPRINT03%20ENGENHARIA%20DE%20DADOS/Notebook%20Vers%C3%A3o%20FINAL%202024-07-12.html)

7 – Conclusão

Foram encontradas limitações na utilização do databricks com uma conta free na plataforma Azure. A importação do arquivo CSV para uma tabela em um banco de dados falhava, pois automaticamente tentava-se criar um cluster com mais de 2 cpus virtuais. A solução encontrada foi criar uma conta Azure de desenvolvedor e *pay as go*.

As sugestões de estudos futuros são a criação de uma modelagem dimensional (modelo estrela) dos dados de segurança pública do estado, propiciando maiores possibilidades de análises e um armazém de dados de melhor performance, projetando relevante crescimento do volume dos dados no futuro e acréscimos de informações sobre a população, mapas e também aumentar a granularidade dos dados com informações por delegacias.

Ficou pendente a implementação do pipeline, pois ainda é necessário entender a frequência e como implementar a automação da aquisição dos dados em questão.

Foi um grande desafio e engrandecedor trabalhar com as novas tecnologias propostas.

Referências:

Site ISP Dados Abertos acessado a partir de <https://www.ispdados.rj.gov.br/>

Site Rio On Watch acessado a partir de <https://rioonwatch.org.br/> e:

[Unidades de Polícia Pacificadora \(UPP\) Parte 1: 2008-2010 \[REFERÊNCIA\] - RioOnWatch](#)

[Unidades de Polícia Pacificadora \(UPP\) Parte 2: 2010-2011 - RioOnWatch](#)

[Unidades de Polícia Pacificadora \(UPP\) Parte 3: 2012 - RioOnWatch](#)

Unidades de Polícia Pacificadora (UPP) Parte 5: 2014 - RioOnWatch