Dados do Atlas do Desenvolvimento Humano no Brasil





Pós-graduação Análise de Big Data



Nome do Aluno:

Nailson dos Santos

Turma:

LAABDDEEAD – Pós Graduação Latu Sensu Especialização em Análise de Big Data – Data Engineering

Repositório do projeto:

https://github.com/nailson-br/tcc_fia



Agenda

- 1. Objetivo do Trabalho
- 2. Proposta da arquitetura
- 3. Ferramentas utilizadas
 - i. Detalhamento das ferramentas utilizadas
- 4. Dicionário de dados
- 5. Análise exploratória (inicial) dos dados
- 6. Considerações finais



Objetivo do Trabalho

Apresentar dados referentes ao **Atlas de Desenvolvimento Humano no Brasil** no **Portal de Dados Abertos do Governo Federal**.

Os dados serão ingeridos em ambiente **Big Data**, serão tratados e distribuídos em bases separadas por assunto, possibilitando o consumo das informações de forma organizada e estruturada.

O resultado final desse trabalho será a disponibilização de **painéis interativos em Power BI** para análise das informações referentes ao IDH.

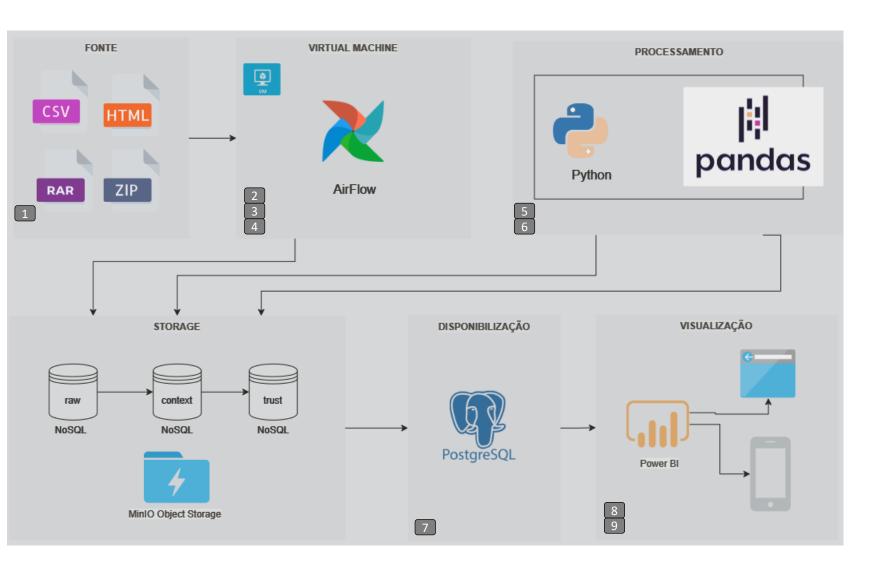
A partir desse painéis, será possível fazer seleção de regiões, UF ou até mesmo, municípios para visualização dos dados por assunto.

Para essa entrega serão disponibilizados os dados referentes aos índices relacionados à saúde.



Arquitetura inicial proposta





- Arquivos disponibilizados no site Portal Brasileiro de Dados Abertos (Governo Federal);
- 2. Obtenção das bases de dados via AirFlow;
- Conversão das fontes em formato padronizado, sem alterações nos dados originais;
- Armazenamento em camada RAW;
- 5. Tratamento dos dados;
- Disponibilização de dados tratados e consolidados em camada CONTEXT;
- 7. Disponibilização de dados em tabelas e/ou views, direcionadas à propósitos e áreas de negócio;
- Consumo das tabelas e geração de relatórios analíticos;
- Disponibilização em navegadores e dispositivos móveis;





Ferramentas utilizadas

O fluxo dos dados, da ingestão à disponibilização, usa as seguintes ferrametentas:

Apache AirFlow

Orquestrador de fluxo de dados para obtenção dos dados externos e ingestão nas camadas do MinIO conforme contexto.

MinIO

Storage de Objetos com divisão nas camadas: RAW – objetos/dados armazenados sem modificação CONTEXT – dados armazenados com tratamento TRUST – dados armazenados já em contexto

PostgreSQL

Banco de dados relacional para camada de disponibilização dos dados.

Power BI

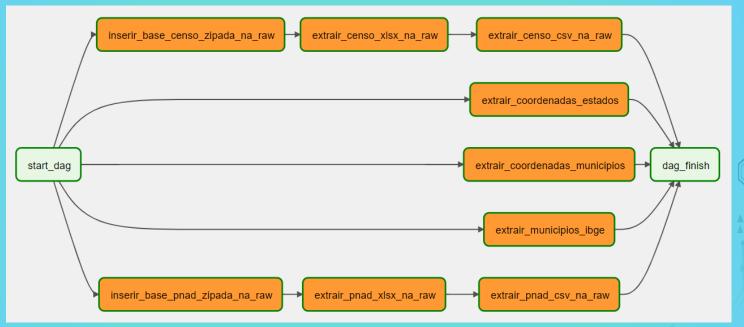
Ferramenta analítica na qual serão exibidos os painéis interativos com os dados disponibilizados.

revent database trace ort("The Rails environment to manning equire 'spec_helper' Modificar imagem equire 'rspec/rails' a seu critério require 'capybara/rspec' require 'capybara/rolls' Capybara.javascript Category.delete_all; Category Shoulda::Matchers.com/spure (8) News config.integrate 🔅 🚾 with.test_fromework program with.library :rolls # Add additional requires below the # Requires supporting name # spec/support/ and its mile # run as spec files by white him # in _spec.rb will have he man # run twice. It is recomm

Detalhamento das ferramentas utilizadas

Apache AirFlow

DAG para ingestão dos dados na RAW



Os dados são obtidos a partir de sites do Governo Federal em formato ZIP. Todos os dados são mantidos: arquivos zip, conteúdo dos arquivos zip e arquivos Excel.

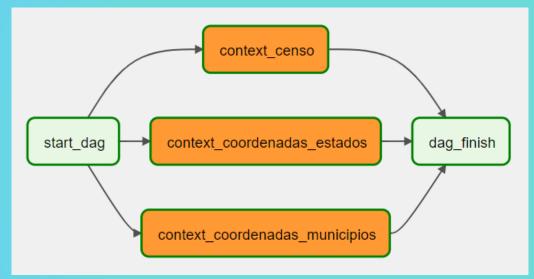
Cada aba dos arquivos Excel são extraídas e gravadas em formato CSV.

revent database ort("The Rails environment to manning equire 'spec_helper' Modificar imagem equire 'rspec/rails" a seu critério require 'capybara/rspec' require 'capybara/roils' Capybara.javascript Category.delete_all; Campa Shoulda::Matchers.com config.integrate & Minim with.test_framework _____ with.library :rails # Add additional regulars because # Requires supporting raw # spec/support/ and its a # run as spec files by warming # in _spec.rb will back by a

Detalhamento das ferramentas utilizadas

Apache AirFlow

DAG para ingestão dos dados na **CONTEXT**



Apenas os dados do censo referente todos os municípios.

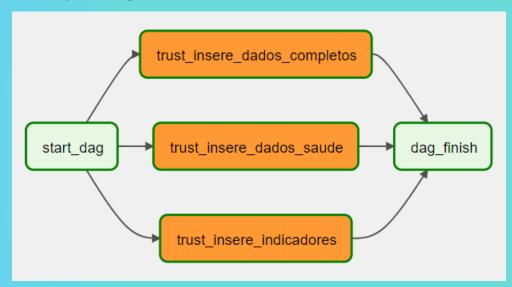
Também estão sendo armazenadas as coordenadas dos municípios a partir de dados do IBGE.

uire File.expand_poth("__/_/see revent database traces ort("The Rails environment is married equire 'spec_helper' Modificar imagem equire 'rspec/rails" a seu critério require 'capybara/rspec' require 'capybara/roils' Capybara.javascript Category.delete_all; Company and Shoulda::Matchers.com/spure (8) News config.integrate 🔅 🚾 with.test_fromework property with.library :rails # Add additional regulars bear as # Requires supporting name and # spec/support/ one its # run as spec files by Amana

Detalhamento das ferramentas utilizadas

Apache AirFlow

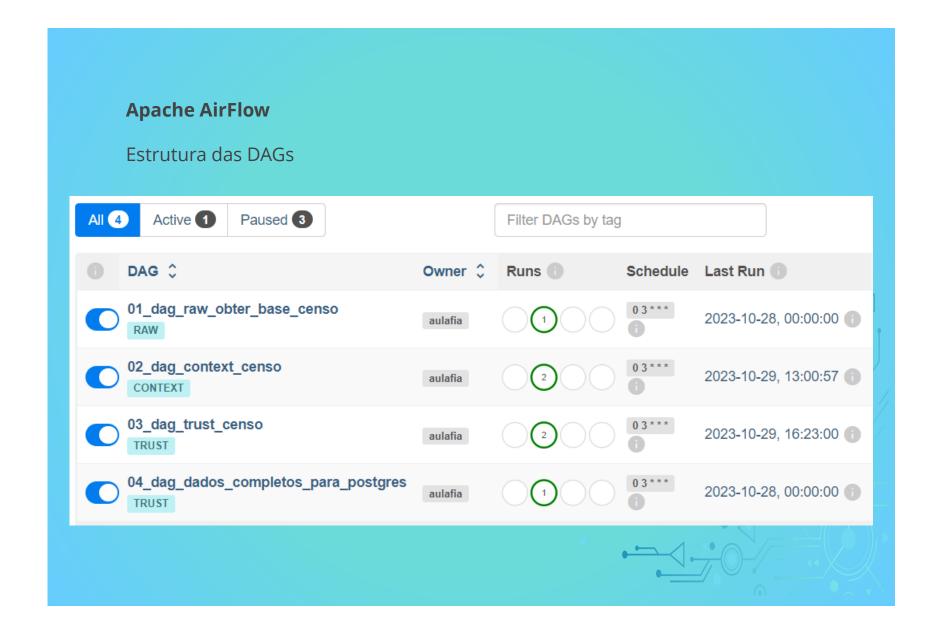
DAG para ingestão dos dados na TRUST



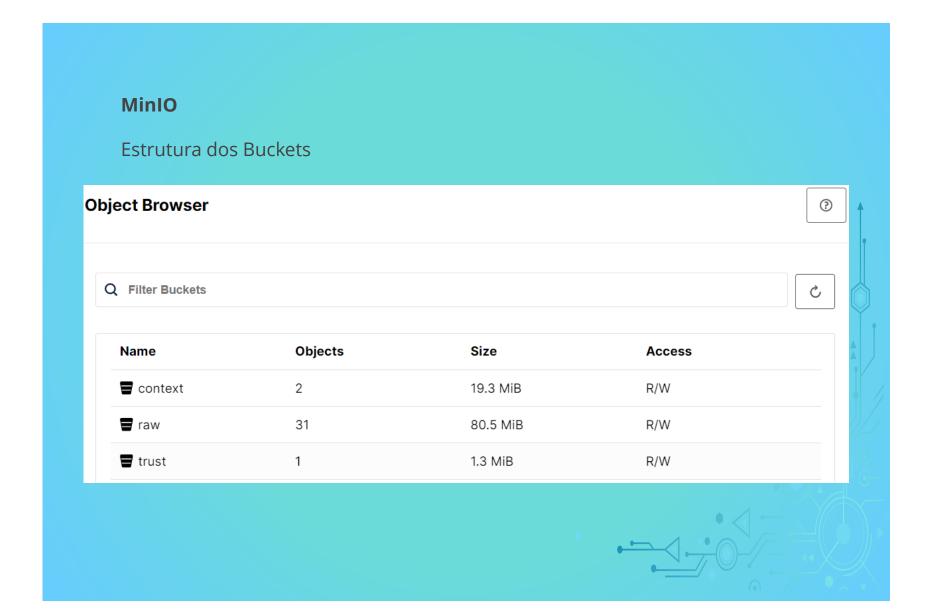
Todos os dados do censo referente todos os municípios.

Também estão sendo armazenadas as coordenadas dos municípios a partir de dados do IBGE.

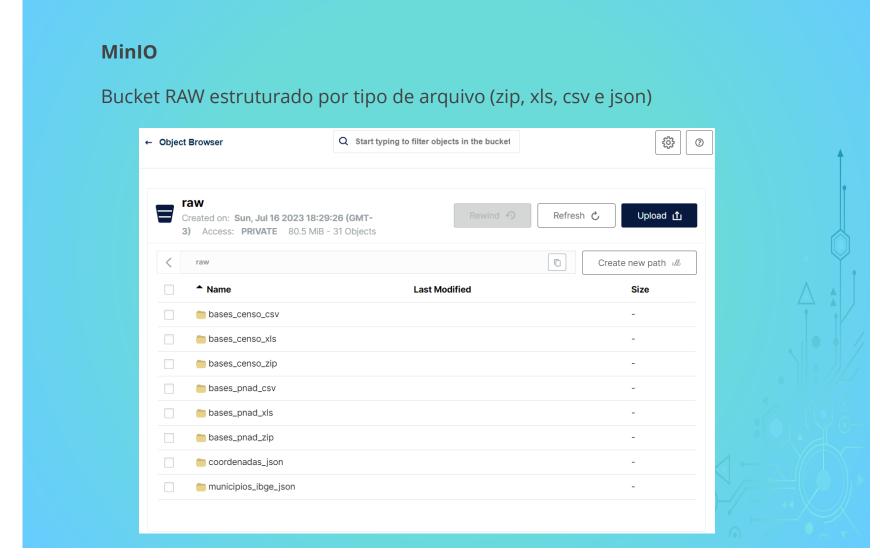
revent database trace ort("The Rails environment to manufacture quire 'spec_helper' Modificar imagem equire 'rspec/rails' a seu critério require 'capybara/rspec' require 'capybara/reils' Capybara.javascript Category.delete_all; Category Shoulda::Matchers.comfigure (a) Manager config.integrate do lateral with.test_framework _____ with.library :reils end # Add additional requires between # Requires supporting name # spec/support/ and its minima # run as spec files by when the # in _spec.rb will back by # run twice. It is recommend



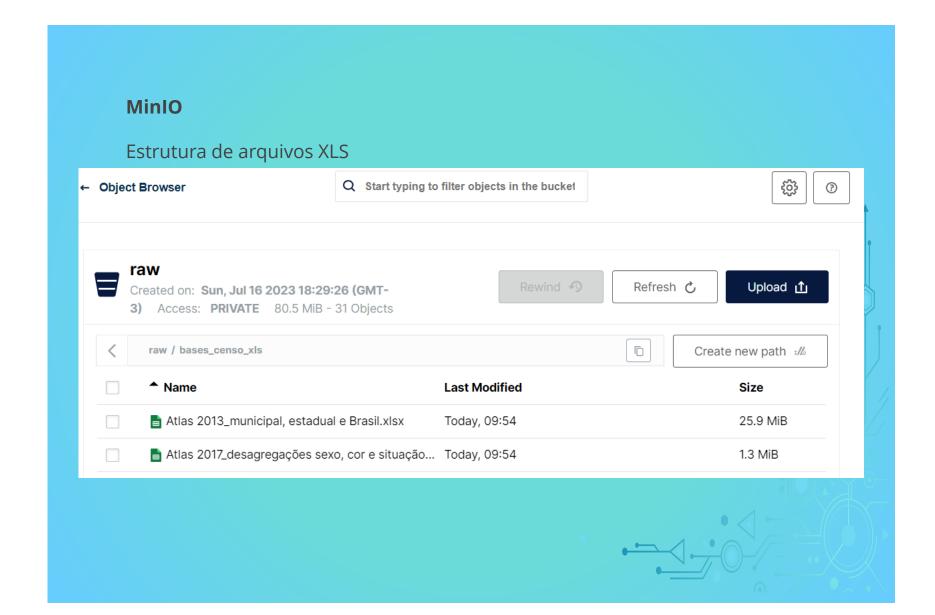
uire File.expand_poth("__/_/see revent database traces ort("The Rails environment is running in equire 'spec_helper' Modificar imagem equire 'rspec/rails' a seu critério require 'capybara/rspec' require 'capybara/rolls' Capybara.javascript Category.delete_all; Campy per Shoulda:: Matchers.comfigure & March config.integrate do lateral with.test_framework present with.library :reils # Add additional regulars bear as 20 # Requires supporting raw # spec/support/ and its == # run as spec files by warming # in _spec.rb will became



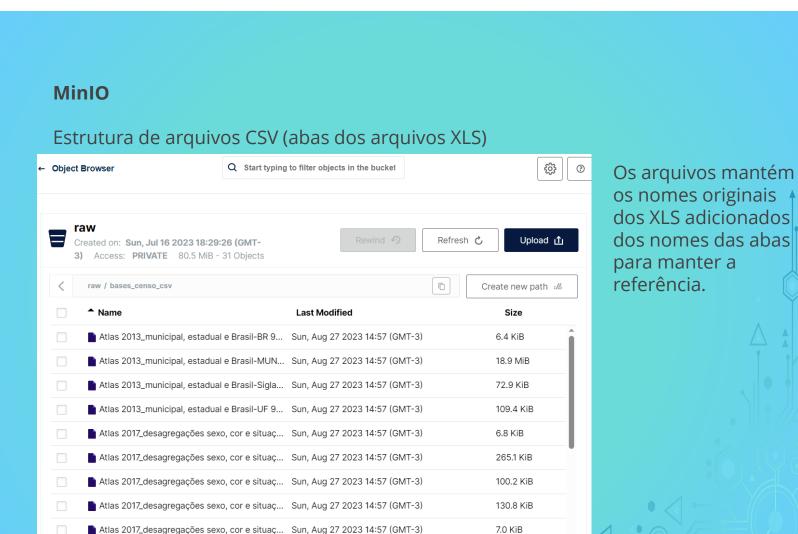
revent database trace ort("The Rails environment is marriage to quire 'spec_helper' Modificar imagem equire 'rspec/rails' a seu critério require 'capybara/rspec' require 'capybara/roils' Capybara.javascript Category.delete_all; (category amount Shoulda::Matchers.comfigure & Names config.integrate & Manual with.test_framework angument with.library :rolls # Add additional requires become # Requires supporting raw # spec/support/ and its summer # run as spec files by when the # in _spec.rb will both by # run twice. It is recommend



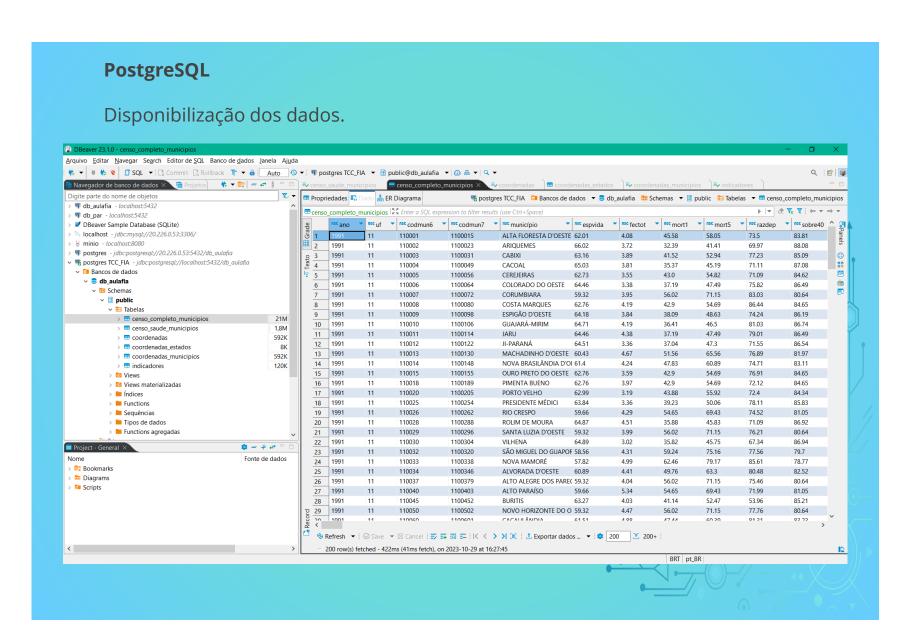
revent database trace ort("The Rails environment is remning the equire 'spec_helper' Modificar imagem equire 'rspec/rails' a seu critério require 'capybara/rspec' require 'capybara/rolls' Capybara.javascript Category.delete_all; Category. Shoulda::Matchers.comfigure (a) Manufacture config.integrate & March with.test_framework _____ with.library :reils end # Add additional requires between # Requires supporting num # spec/support/ and its summan # run as spec files by when the # in _spec.rb will back by # run twice. It is recommend



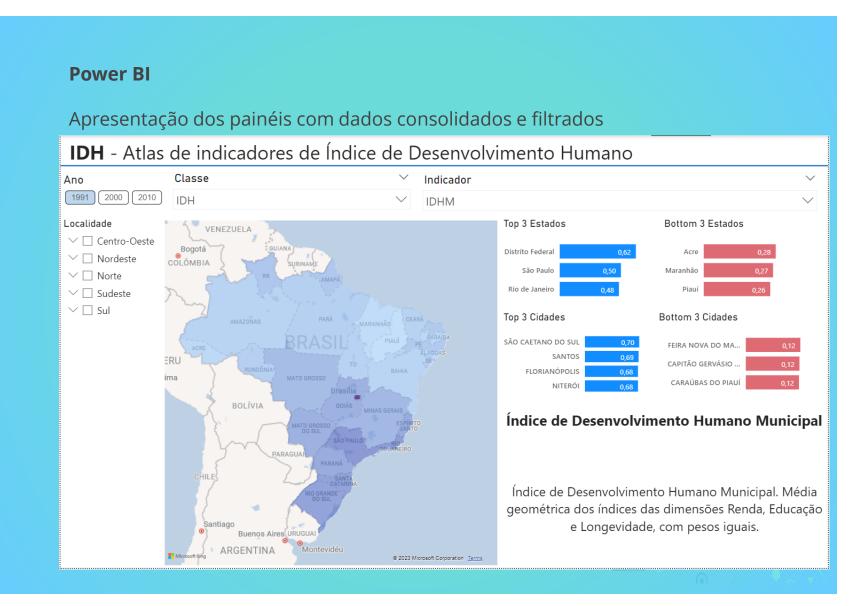
uire File.expand_path(". revent database ort("The Rails environment to manufacture quire 'spec_helper' Modificar imagem equire 'rspec/rails' a seu critério require 'capybara/rspec' require 'capybara/reils' Capybara.javascript Category.delete_all; Category Shoulda:: Matchers.comfigure & Manager config.integrate do lateral with.test_framework _____ with.library :rails # Add additional requires between # Requires supporting the # run as spec files by domining the # spec/support/ and its # in _spec.rb will have but among # run twice. It is recommun



revent database traces ort("The Rails environment to manage to equire 'spec_helper' Modificar imagem equire 'rspec/rails' a seu critério require 'capybara/rspec' require 'capybara/reils' Capybara.javascript Category.delete_all; Company Shoulda::Matchers.comfigure & New York config.integrate do lateral with.test_framework with.library :rails end # Add additional regains have # Requires supporting # spec/support/ # run as spec files # in _spec.rb will # run twice. It is # end with _spec. # end with _spec.



wire File.expand_path("..." revent database trace ort("The Rails environment to m quire 'spec_helper' Modificar imagem equire 'rspec/rails' a seu critério require 'capybara/rspec' require 'capybara/rolls' Capybara.javascript Category.delete_all; Category and Shoulda::Matchers.com config.integrate & MANN with.test_framework _____ with.library :rails # Add additional requires being the # Requires supporting name and # spec/support/ and its a-# run as spec files by warming # in _spec.rb will back by and # run twice. It is recomm





Definição das camadas do DataLake

Camada RAW

Nessa camada são armazenados os dados originais, obtidos do site do governo federal, em formato ZIP.

Esse bucket está estruturado para separar formatos zip, xls, csv e json.

O conteúdo de cada arquivo ZIP descompactado e armazenado na RAW sem tratamento.

Abas de arquivos XLS são extraídas e armazenadas sem tratamento, em formato CSV.



Definição das camadas do DataLake

Camada CONTEXT

Nessa camada são armazenados os dados originais, porém já filtrados para o conceito dos índices de saúde que compõem o IDH.

Essa etapa está sendo tratada como PoC, portanto está sendo testada a arquitetura proposta e os mecanismos de ingestão.

Paras as próximas etapas serão também trazidos para essa camada os dados referentes a outros índices que compõem o IDH.



Definição das camadas do DataLake

Camada TRUST

Nessa camada são armazenados os dados referentes aos censos de 1991, 2000 e 2010

Nessa camada foram adicionados dados de classificação dos índices levantados pelos censos nacionais.

Inseridos também dados do IBGE com informações de localização e nomes das entidades federais (estados e municípios).



Dicionário de dados

Base de dados do censo com os índices de saúde

SIGLA	NOME CURTO	NOME LONGO	DEFINIÇÃO
UF	Código da Unidade da Federação	Código da Unidade da Federação	Código utilizado pelo IBGE para identificação do estado.
UFN	Nome da Unidade da Federação	Nome da Unidade da Federação	Nome da Unidade da Federação.
CODMUN6	Código do Município	Código do Município	Código utilizado pelo IBGE para identificação do município.
CODMUN7	Código do Município	Código do Município	Código utilizado pelo IBGE para identificação do município(com digito verificador).
NOMEMUN	Nome do Município	Nome do município	Nome do município.
ESPVIDA	Esperança de vida ao nascer	Esperança de vida ao nascer	Número médio de anos que as pessoas deverão viver a partir do nascimento, se permanecerem constantes ao longo da vida o nível e o padrão de mortalidade por idade prevalecentes no ano do Censo.
FECTOT	Taxa de fecundidade total	Taxa de fecundidade total	Número médio de filhos que uma mulher deverá ter ao terminar o período reprodutivo (15 a 49 anos de idade).
MORT1	Mortalidade infantil	Mortalidade até um ano de idade	Número de crianças que não deverão sobreviver ao primeiro ano de vida em cada 1000 crianças nascidas vivas.
MORT5	Mortalidade até 5 anos de idade	Mortalidade até cinco anos de idade	Probabilidade de morrer entre o nascimento e a idade exata de 5 anos, por 1000 crianças nascidas vivas.
RAZDEP	Razão de dependência	Percentual da população de menos de 15 anos e da população de 65 anos e mais em relação à população de 15 a 64 anos	Razão de dependência é medida pela razão entre o número de pessoas com 14 anos ou menos e de 65 anos ou mais de idade (população dependente) e o número de pessoas com idade de 15 a 64 anos (população potencialmente ativa) multiplicado por 100.
SOBRE40	Probabilidade de sobrevivência até 40 anos	Probabilidade de sobrevivência até 40 anos	Probabilidade de uma criança recém-nascida viver até os 40 anos, se permanecerem constantes ao longo da vida o nível e o padrão de mortalidade por idade prevalecentes no ano do Censo.
SOBRE60	Probabilidade de sobrevivência até 60 anos	Probabilidade de sobrevivência até 60 anos	Probabilidade de uma criança recém-nascida viver até os 60 anos, se permanecerem constantes ao longo da vida o nível e o padrão de mortalidade por idade prevalecentes no ano do Censo.
T_ENV	Taxa de envelhecimento	Taxa de envelhecimento	Razão entre a população de 65 anos ou mais de idade e a população total multiplicado por 100.



Dicionário de dados

Base de dados do IBGE com as coordenadas dos municípios

nome	descrição
codigo_ibge	Código do Município
nome	Nome do município
latitude	Coordenada de latitude
longitude	Coordenada de longitude
capital	Código da capital
codigo_uf	Código da Unidade da Federação
siafi_id	Código registrado no SIAFI
ddd	Código de discagem à distância
fuso_horario	fuso horário



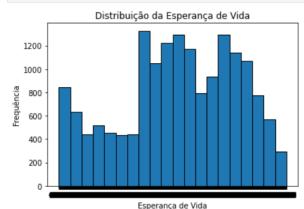
Análise exploratória dos dados

Análise dos dados da tabela de dados de saúde

```
[8]: print(df.describe())
                       UF Codmun6 Codmun7
                                               Município ESPVIDA FECTOT
             16695
                            16695
                                     16695
                                                   16695
                                                                        16695
     count
                             5565
                                      5565
     unique
     top
              1991
                       31
                          110001
                                  1100015
                                            SÃO DOMINGOS
                     2559
                                                      15
                                                                   147
                                                                           71
     freq
             MORT5 RAZDEP SOBRE40 SOBRE60
                                          T_ENV
     count
     unique
                             1883
                                            1327
              40.6 53.35
                                   76.29
                                            5.63
                            92.11
     top
                       12
                               55
                                       44
     frea
```

```
[10]: import matplotlib.pyplot as plt

plt.hist(df['ESPVIDA'], bins=20, edgecolor='k')
plt.xlabel('Esperança de Vida')
plt.ylabel('Frequência')
plt.title('Distribuição da Esperança de Vida')
plt.show()
```



```
plt.bar(df['UF'], df['T_ENV'])
plt.xlabel('Unidade Federativa')
plt.ylabel('Taxa de Envelhecimento')
plt.title('Taxa de Envelhecimento por Unidade Federativa')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

Taxa de Envelhecimento por Unidade Federativa

Unidade Federativa

Considerações finais



Finalização da PoC

Essa entrega, a terceira entrega do trabalho, encerra o trabalho apresentando arquitetura, todas as etapas de ingestão de dados, disponibilização nas camadas do repositório e exibição das informações de forma tabulada em um painel interativo feito em Power BI.

Como resultado observa-se uma solução capaz de adquirir dados e fazer a ingestão em uma estrutura de armazenamento com facilidades de diversidade de formatos, segurança e escalabilidade.

Os dados são distribuídos em camadas que permitem a consulta e montagem de novos painéis e conjunto de dados, pois foram observados todos os critérios para manter os dados originais intactos.

Estão disponíveis no repositório dados do CENSO e do PNAD, permitindo novas consultas e painéis.

Repositório do projeto

https://github.com/nailson-br/tcc fia