Aluno: Alessandro Jatobá

MVP - Engenharia de dados - 2024

Objetivo

A Organização Mundial de saúde (OMS) define que sistemas de saúde resilientes devem desenvolver um conjunto de atributos definidos a partir de quatro dimensões: Força de Trabalho, Provimento, Financiamento e Governança. Cada uma dessas dimensões é definida a partir de indicadores relacionados à manutenção de funções essenciais de saúde pública.

O objetivo desse MVP é desenvolver um Data Warehouse que permita visualizar o desenvolvimento dessas dimensões na última década, permitindo identificar seu comportamento em determinados momentos, como naqueles de crise de financiamento ou de involução da força de trabalho, nas principais capitais brasileiras (a partir de indicadores como o Índice de Desenvolvimento Humano (IDH) e População)

Questões a serem respondidas

- Qual a média de cada dimensão nos últimos 10 anos?
- Qual dimensão apresentou o maior desenvolvimento ao longo dos últimos 10 anos?
- No ano em que a dimensão de Financiamento apresentou sua pior baixa dos últimos 10 anos, qual foi o comportamento das dimensões Força de Trabalho e Provimento?

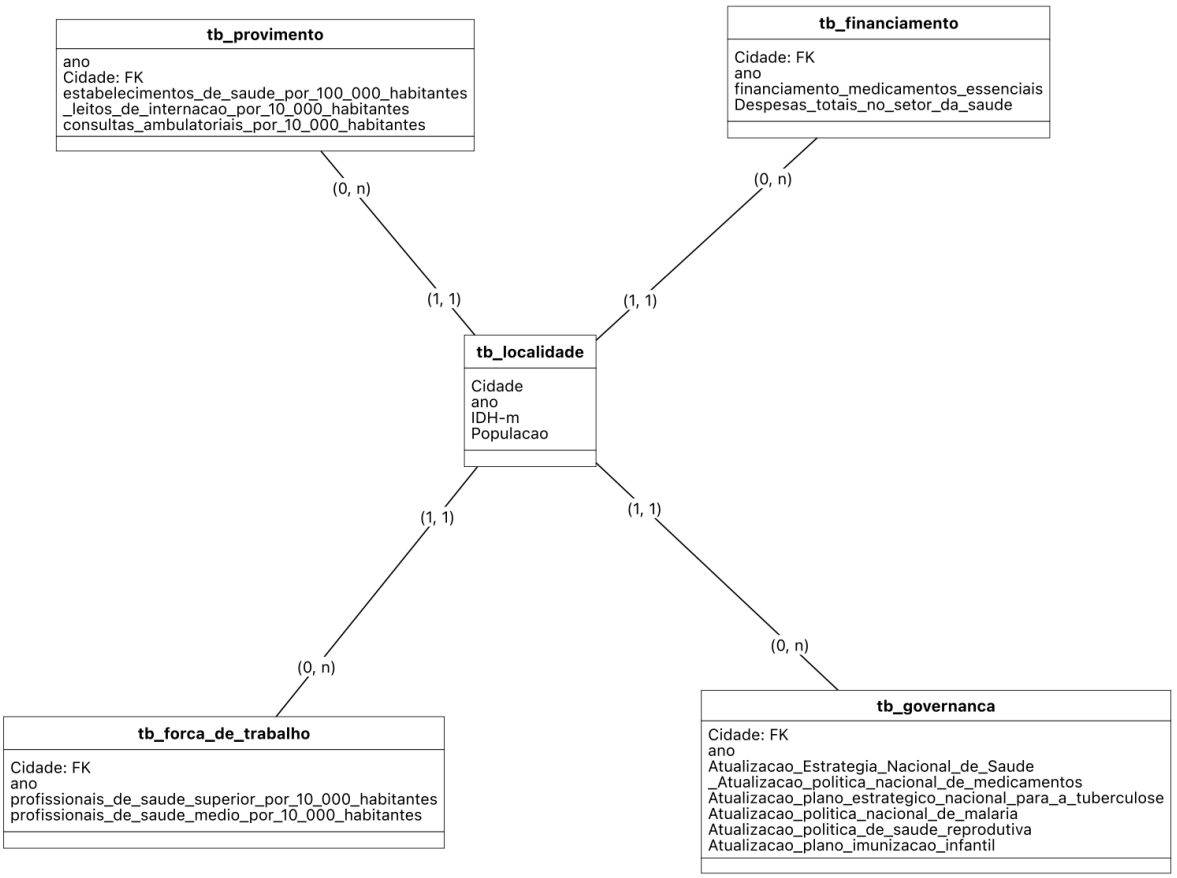
Busca pelos dados

Coleta

Todos os dados utilizados são de acesso público e disponíveis no barramento de dados do Departamento de Informática do Ministério da Saúde (DATASUS). Os seguintes sistemas serviram como fonte de dados:

- Sistema de Informações Hospitalares (SIA)
- Sistema de Informações Ambulatoriais (SIH)
- Sistema Nacional de Doenças e Agravos de Notificação (SINAN)
- Sistema de Informações da Atenção Básica (SISAB)
- Cadastro Nacional de Estabelecimentos de Saúde (CNES)

Modelagem



Catálogo de dados

tb_localidade
Descrição: armazena as capitais analisadas e suas características gerais
Fonte de dados: SISAB, dados públicos IBGE, Censo 2022
Campos:
ano

- Tipo: Numérico inteiro
- Mínimo: 2010
- Máximo: 2024
- Comentário: contém os anos a que cada variável da localidade se refere

Cidade

- Tipo: Caracter
- Comentário: Nomes das cidades

IDH-m;

- Tipo: Numérico de ponto flutuante
- Mínimo: 0
- Máximo: 1
- Comentário: IDH médio das cidades

Populacao

- Tipo: Numérico inteiro
- Mínimo: 0
- Máximo: N/A
- Comentário: População das cidades

tb_forca_de_trabalho
Descrição: armazena dados sobre os profissionais de saúde atuando no país
Fonte de dados: CNES
Campos:
ano

- Tipo: Numérico inteiro
- Mínimo: 2010
- Máximo: 2024
- Comentário: contém os anos a que cada variável da localidade se refere

Cidade

- Tipo: Caracter
- Comentário: Cidade (Chave estrangeira com origem na tabela tb_localidade)

profissionais_de_saude_superior_por_10_000_habitantes

- Tipo: Numérico inteiro
- Mínimo: 0
- Máximo: N/A
- Comentário: Quantidade de profissionais de saúde de nível superior cadastrados para a cidade em questão no CNES

profissionais_de_saude_medio_por_10_000_habitantes

- Tipo: Numérico inteiro
- Mínimo: 0
- Máximo: N/A
- Comentário: Quantidade de profissionais de saúde de nível médio cadastrados para a cidade em questão no CNES

tb_provimento
Descrição: armazena dados sobre a utilização de serviços de saúde no país
Fonte de dados: CNES, SIA e SIH
Campos:
ano

- Tipo: Numérico inteiro
- Mínimo: 2010
- Máximo: 2024
- Comentário: contém os anos a que cada variável da localidade se refere

Cidade

- Tipo: Caracter
- Comentário: Cidade (Chave estrangeira com origem na tabela tb_localidade)

estabelecimentos_de_saude_por_100_000_habitantes

- Tipo: Numérico inteiro
- Mínimo: 0
- Máximo: N/A
- Comentário: Quantidade de unidades de saúde cadastrados para a cidade em questão no CNES_

leitos_de_internacao_por_10_000_habitantes;

- Tipo: Numérico inteiro
- Mínimo: 0
- Máximo: N/A
- Comentário: Quantidade de leitos de internação cadastrados para a cidade em questão no SIH

consultas_ambulatoriais_por_10_000_habitantes

- Tipo: Numérico inteiro
- Mínimo: 0
- Máximo: N/A
- Comentário: Quantidade de consultas ambulatoriais realizadas para a cidade em questão cadastradas no sistema SIA

tb_financiamento
Descrição: armazena dados sobre as despesas com serviços de saúde no país
Fonte de dados: SINAN
Campos:
ano

- Tipo: Numérico inteiro
- Mínimo: 2010
- Máximo: 2024
- Comentário: contém os anos a que cada variável da localidade se refere

Cidade

- Tipo: Caracter
- Comentário: Cidade (Chave estrangeira com origem na tabela tb_localidade)

financiamento_medicamentos_essenciais

- Tipo: Numérico de ponto flutuante
- Mínimo: 0
- Máximo: N/A
- Comentário: Valores gastos na aquisição de medicamentos essenciais, conforme cadastrado no sistema SINAN

Despesas_totais_no_setor_da_saude

- Tipo: Numérico de ponto flutuante
- Mínimo: 0
- Máximo: N/A
- Comentário: Valores gastos na aquisição de medicamentos essenciais, conforme cadastrado no sistema SINAN

tb_governancsa
Descrição: armazena dados sobre a implementação de políticas de saúde no país
Fonte de dados: SISAB, CNES
Campos:
ano
- Tipo: Numérico inteiro
- Mínimo: 2010
- Máximo: 2024
- Comentário: contém os anos a que cada variável da localidade se refere

Cidade
- Tipo: Caracter
- Comentário: Cidade (Chave estrangeira com origem na tabela tb_localidade)

Atualizacao_Estrategia_Nacional_de_Saude
- Tipo: Booleano (Sim/Não)
- Comentário: A cidade realizou atualização atualização da Estratégia Nacional de saúde no ano em questão

Atualizacao_politica_nacional_de_medicamentos;
- Tipo: Booleano (Sim/Não)
- Comentário: A cidade realizou atualização da Política Nacional de Medicamentos no ano em questão

Atualizacao_plano_estrategico_nacional_para_a_tuberculose
- Tipo: Booleano (Sim/Não)
- Comentário: A cidade realizou atualização da Estratégia Nacional de Tuberculose no ano em questão

Atualizacao_politica_nacional_de_malaria
- Tipo: Booleano (Sim/Não)
- Comentário: A cidade realizou atualização da Política Nacional de Malária no ano em questão

Atualizacao_politica_de_saude_reprodutiva
- Tipo: Booleano (Sim/Não)
- Comentário: A cidade realizou atualização da Política Nacional de Saúde Reprodutiva no ano em questão

Atualizacao_plano_imunizacao_infantil
- Tipo: Booleano (Sim/Não)
- Comentário: A cidade realizou atualização do Plano de Imunização Infantil no ano em questão

Criação do Ambiente de Data Warehouse

Foi utilizada a plataforma Databricks para a hospedagem do Datawarehouse proposta neste MVP, como mostra a figura 1. O Pipeline é intitulado "Resiliência dos Sistemas de Saúde"

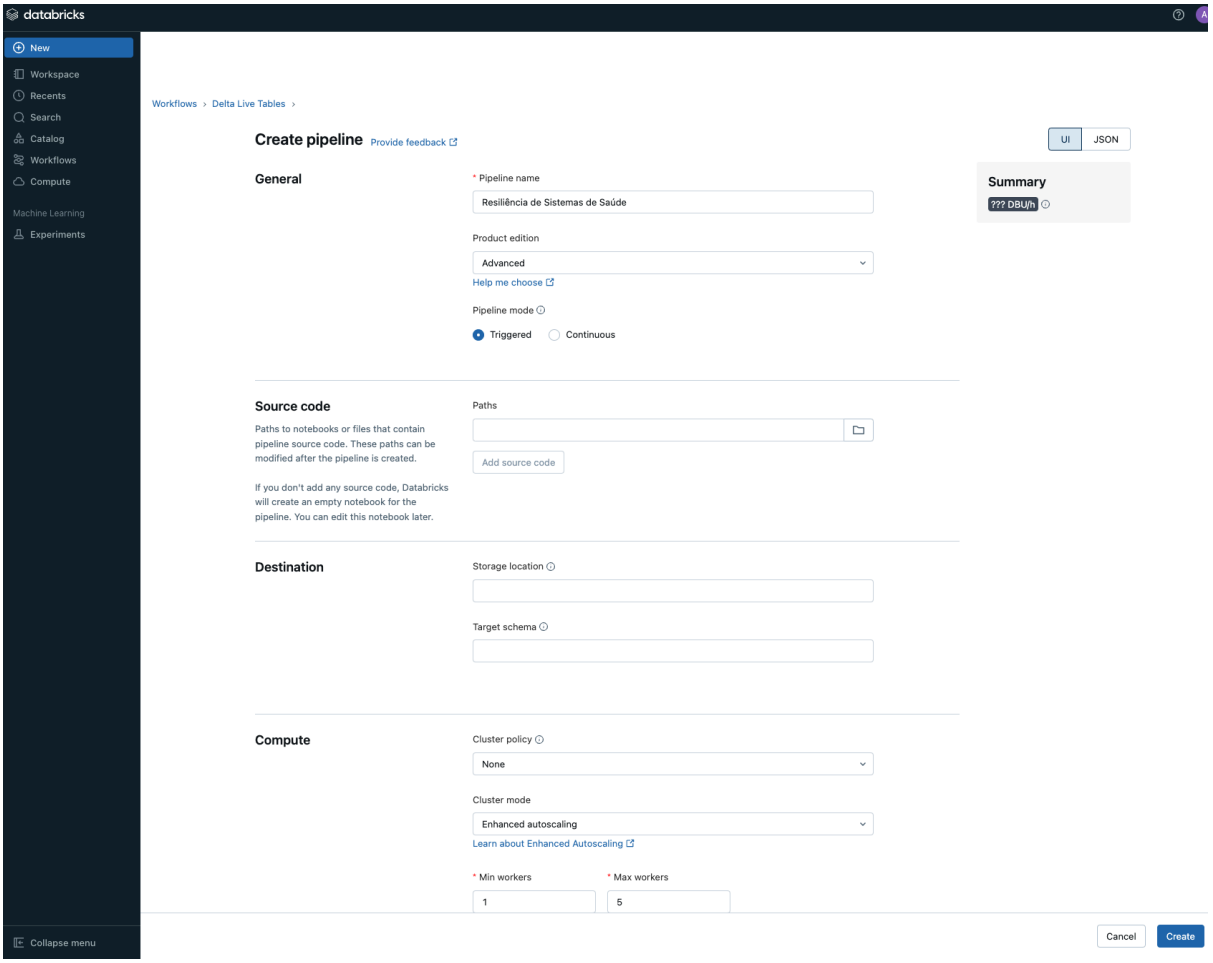


Figura 1: criação do pipeline no Databricks

Criando o Cluster

O cluster utilizado nesse projeto foi criado a partir dos serviços de nuvem do Google Cloud. Depois do período de inatividade de 60 minutos, é necessário criar novo cluster no Databricks (uma vez que foi utilizada a versão *Community* da plataforma)

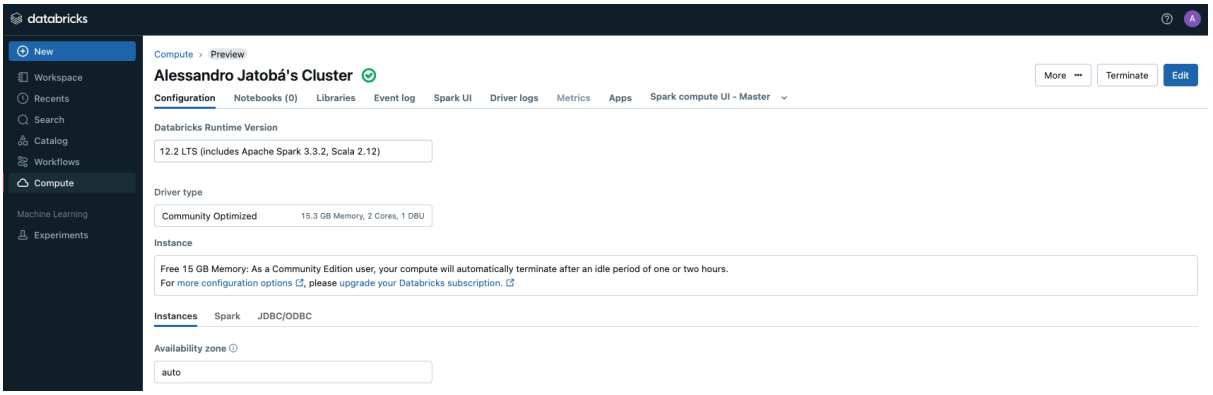


Figura 2: Criação e inicialização do cluster

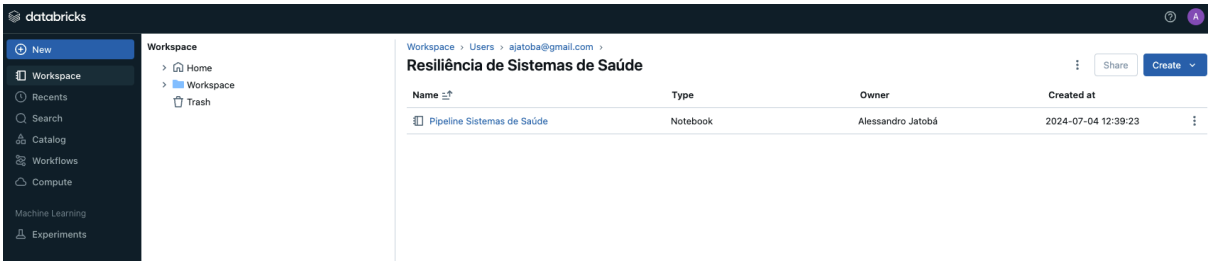


Figura 3: Visão geral do workspace do Databricks com o pipeline

Análise

Como forma de demonstrar melhor as relações entre as dimensões, aspecto essencial das questões a serem respondidas, foram escolhidos diferentes tipos de visualizações de dados. Essa também foi uma interessante estratégia de aprendizado, colocando em perspectiva a construção de dashboards interessantes. Abaixo é possível ver os gráficos elaborados:

```
    ax.fill(angles, values, 'b', alpha=0.1)

plt.legend(loc='upper right', bbox_to_anchor=(0.1, 0.1))
plt.show()
```

(7) Spark Jobs

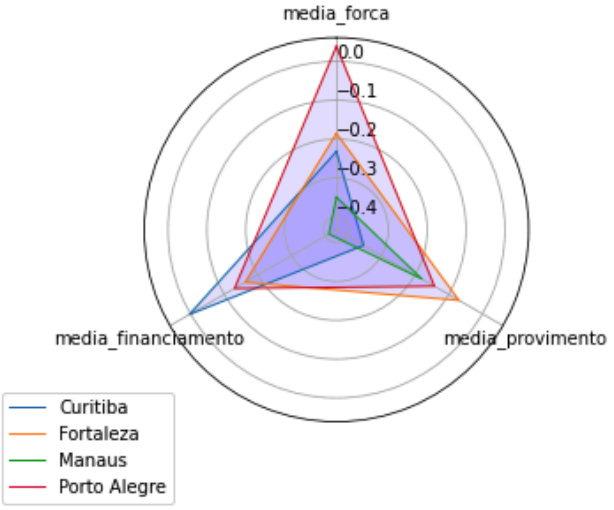


Figura 4: Gráfico de radar

```
plt.plot(ano, media_financiamento, label='Financiamento')

plt.legend()
plt.show()
```

(27) Spark Jobs

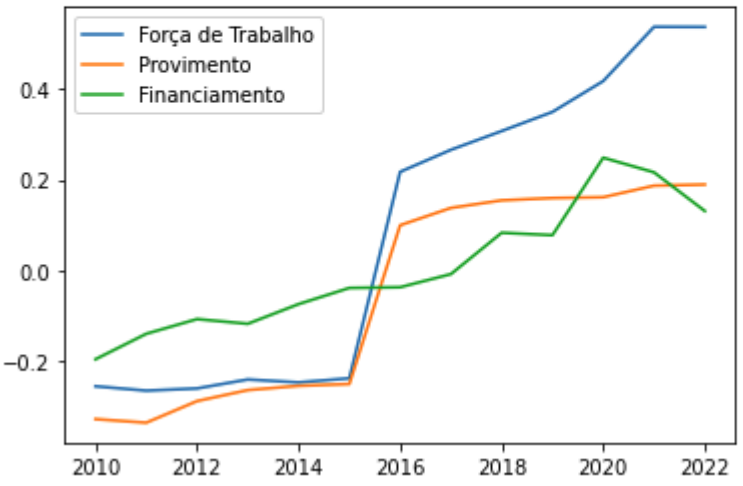


Figura 5: Gráfico de linhas

```
22   plt.legend()
23   plt.show()
```

(22) Spark Jobs

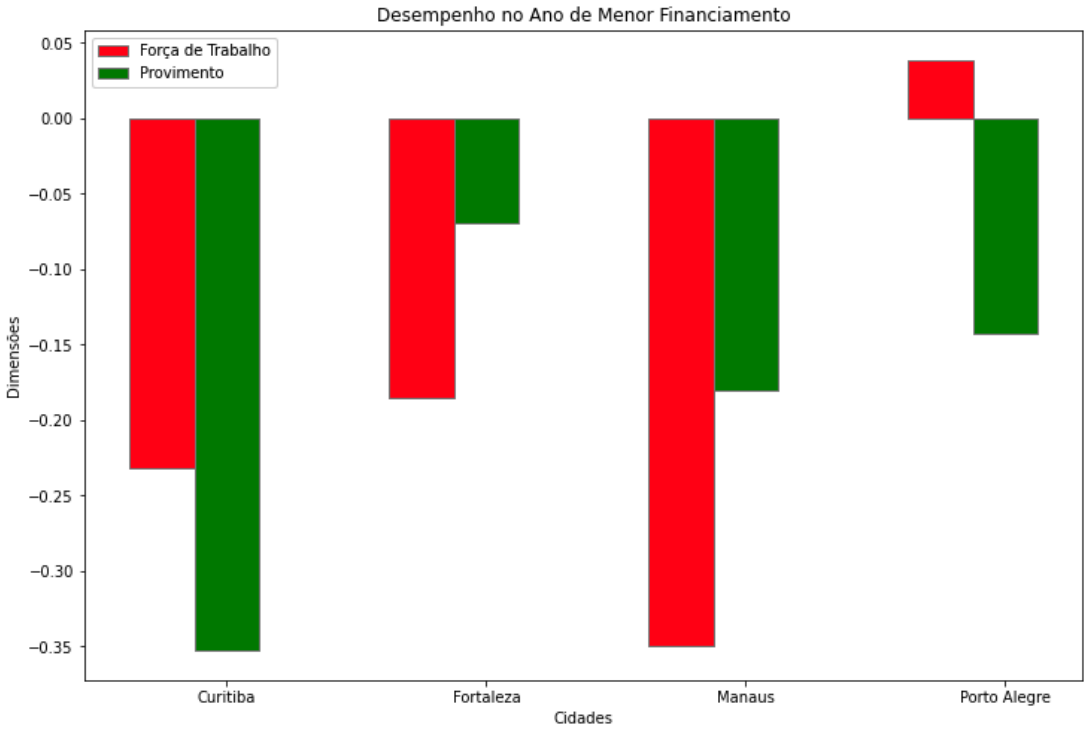


Figura 6: Gráfico de barras

A figura 7 mostra uma visão geral do catálogo de dados do pipeline criado nesse MVP no Databricks

Figura 7: Visão geral do catálogo de dados com as tabelas do pipeline

Autoavaliação

Nem todos os objetivos previstos inicialmente puderam ser atingidos. Especialmente, dada a granularidade dos dados, outras análises poderiam ser feitas. Optou-se, portanto, em fazer análises de mais alto nível, utilizando valores médios entre as variáveis de cada dimensão.

O nível de dificuldade para a elaboração das queries usando SQL foi considerado alto, talvez devida a minha pouca familiaridade com a disciplina. De qualquer forma, foi proveitoso e fui capaz de elaborar queries de complexidade média, envolvendo JOIN e funções SQL como média (AVG) e Máximo/Mínimo (MAX/MIN). Foi bastante interessante poder realizar pesquisas para descobrir a maneira de elaborar essas consultas, o que me permitiu aprender bastante.