

1. Objetivo

Meu objetivo é analisar dados relacionados aos casos nacionais de COVID-19, com intuito de saber sobre o número de contaminações confirmados e mortes ocasionados pela pandemia nos estados brasileiros, a fim de responder as seguintes perguntas:

- Quantos casos de Covid-19 confirmados?
- Quantas mortes por causa da Covid?
- Qual é o estado com maior volume de casos de Covid confirmados?
- Qual é o estado com maior número de mortes?
- Qual estado com o maior índice de morte em relação aos casos confirmados?
- Qual estado com o menor índice de morte em relação aos casos confirmados?
- Qual estado com o menor índice de contágio para cada 100 mil habitantes?

2. Dataset

Dataset Covid-19: arquivo caso.csv.gz

Boletins informativos e casos do coronavírus por município por dia

- **Fonte original:** Secretarias de Saúde estaduais
- **Libertado por:** Álvaro Justen e dezenas de colaboradores/Brasil.IO
- **Código-fonte:** <https://github.com/turicas/covid19-br>
- **Licença:** Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)
- **Links relacionados:** Boletins PR, Boletins SP, Boletins RO, Boletins MG, Boletins RS, Boletins MT, Boletins MS, Boletins BA, Boletins PE, Informações sobre a coleta de dados (manual), Boletins AC, Boletins AL, Boletins AM, Boletins AP, Boletins CE, Boletim ES, Boletins GO, Boletins MA, Boletins PA, Boletins PB, Boletins PI, Boletins RJ, Boletins RN, Boletins RR, Boletins SC (1), Boletins SC (2), Boletins DF (1), Boletins DF (2), Boletins SE, Boletins TO, Boletins RJ (2), Documentação da API, Perguntas e respostas sobre os dados, Portal da Transparência do Registro Civil



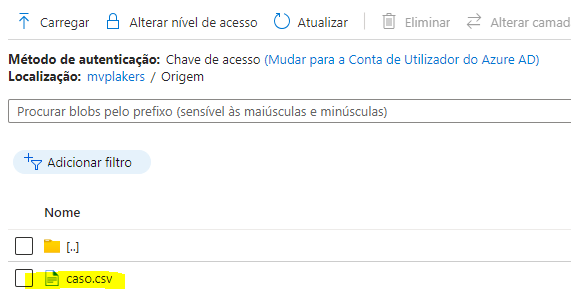
- **date:** data de coleta dos dados no formato YYYY-MM-DD.
- **state:** sigla da unidade federativa, exemplo: SP.
- **city:** nome do município (pode estar em branco quando o registro é referente ao estado, pode ser preenchido com Importados/Indefinidos também).
- **place_type:** tipo de local que esse registro descreve, pode ser city ou state.

- **order_for_place**: número que identifica a ordem do registro para este local. O registro referente ao primeiro boletim em que esse local aparecer será contabilizado como 1 e os demais boletins incrementarão esse valor.
- **is_last**: campo pré-computado que diz se esse registro é o mais novo para esse local, pode ser True ou False.
- **city_ibge_code**: código IBGE do local.
- **confirmed**: número de casos confirmados.
- **deaths**: número de mortes.
- **estimated_population**: população estimada para esse município/estado em 2020
- **estimated_population_2019**: população estimada para esse município/estado em 2019, segundo o IBGE. Essa coluna possui valores desatualizados
- **confirmed_per_100k_inhabitants**: número de casos confirmados por 100.000 habitantes (baseado em estimated_population).
- **death_rate**: taxa de mortalidade (mortes / confirmados).

3. Coleta dos dados

Nesta etapa começo o processo da coleta dos dados, onde faço download do arquivo caso.csv.gz do site Brasil.IO para o meu Desktop. Após este passo, descomprimo o arquivo .gz, extraindo o arquivo "caso.csv". Feito isso, prossigo fazendo o upload do arquivo da minha máquina para o armazenamento do Azure.

3.1. Criando conta de armazenamento



A partir deste momento usarei a ferramenta Azure Synapse Analytics da Microsoft, uma plataforma que oferece vários mecanismos de análise para ajudar você a ingerir, transformar, modelar e analisar seus dados.

3.2. Etapas de criação da área de trabalho do Synapse Analytics

Criar área de trabalho do Synapse ...

*** Informações básicas** * Segurança Rede Etiquetas Rever + criar

Crie uma área de trabalho do Synapse para desenvolver uma solução de análise empresarial em apenas alguns cliques.

Detalhes do projeto

Selecione a assinatura para gerir os recursos implementados e custos. Utilize os grupos de recursos como pastas para organizar e gerir todos os recursos.

Assinatura *

Grupo de recursos * [Criar novo](#)

Grupo de recursos geridos ☒

Detalhes da área de trabalho

Dê um nome à sua área de trabalho, selecione um local e escolha um sistema de ficheiros do Data Lake Storage Gen2 primário para servir de local predefinido para registos e saída de trabalho.

Nome da área de trabalho * ☒

Região *

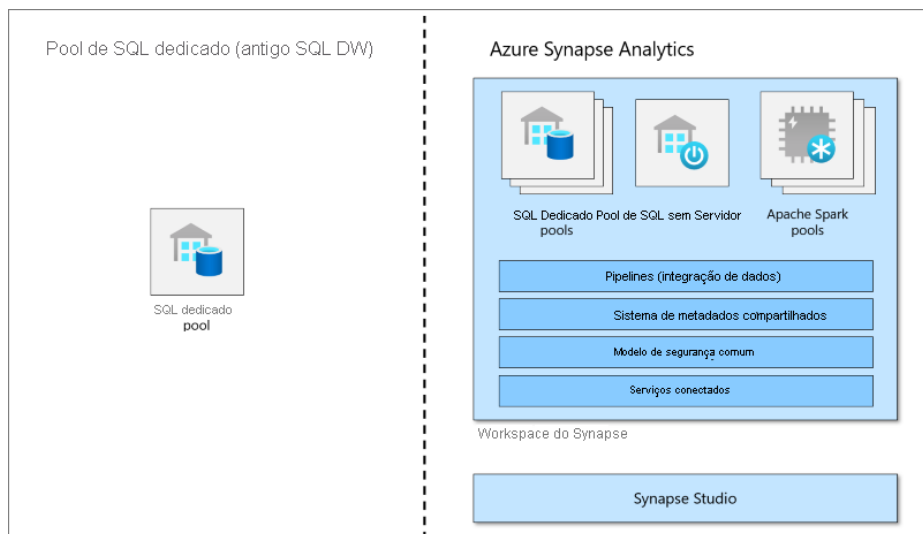
Selecione Data Lake Storage Gen2 * ☒ Da assinatura ☐ Manualmente através de URL

Nome da conta * [Criar novo](#)

Nome do sistema de ficheiros * [Criar novo](#)

3.3. Pool de SQL dedicado

O pool de SQL dedicado (antigo SQL DW) refere-se aos recursos de data warehouse empresariais que estão disponíveis no Azure Synapse Analytics.



3.3.1.1. Etapas de criação do Pool

Novo conjunto de SQL dedicado ...

* Informações básicas * Definições adicionais Etiquetas Rever + criar

Crie um conjunto de SQL dedicado com as suas configurações preferidas. Preencha o separador Informações Básicas e, em seguida, aceda a Rever + Criar para aprovisionar com as predefinições inteligentes ou visite cada separador para personalizar. [Mais informações](#)

Detalhes do conjunto de SQL dedicado

Atribua um nome ao conjunto de SQL dedicado e escolha as definições iniciais.

Nome do conjunto de SQL dedicado * ✓

Georredundante * ☐ Sim ☒ Não

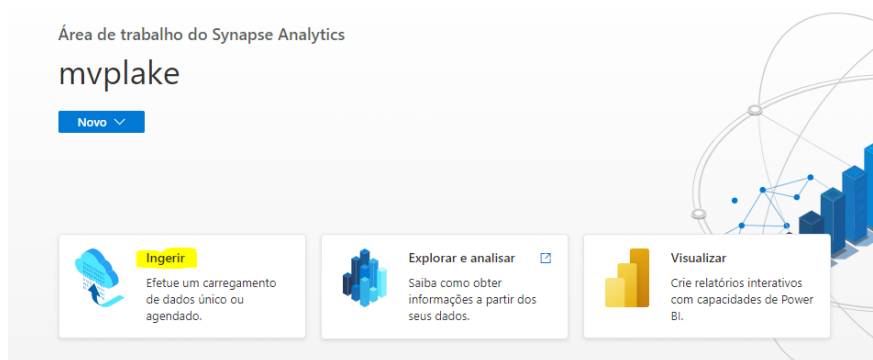
⚠ Os seus dados do conjunto de SQL dedicado do Azure Synapse Analytics não serão replicados numa [região emparelhada](#) para proteção contra indisponibilidades. Esta opção garante que os seus dados não saem do limite geográfico do seu país. Se quiser ativar a replicação na sua [região emparelhada](#) para proteger os dados, seleccione "Sim".

Nível de desempenho ☐ DW100c

Preço estimado ☐ **Custo Estimado por Hora**
2.42 USD
[Ver detalhes da preços](#)

4. Ingestão dos dados

A Ingestão de Dados pode ser definida como sendo o processo usado na absorção de dados de uma grande variedade de fontes, fazendo em seguida a sua transferência para determinado destino onde então serão finalmente analisados.




Nesta etapa pretendo copiar as linhas do arquivo "caso.csv", que está dentro do storage do Azure, para uma tabela de nome "dbo.caso", inicialmente sem nenhum tratamento dos dados para uma pré-análise.


4.1. Ferramenta Copiar Dados

Esta ferramenta cria um pipeline único ou programado para carregar dados de mais de 90 fontes de dados.

Tipo de tarefa

**Tarefa de cópia incorporada**

Obterá um único pipeline para copiar dados facilmente a partir de mais de 90 fontes de dados.

**Tarefa de cópia orientada por metadados**

Obterá pipelines parametrizados que podem ler metadados de um arquivo externo para carregar dados em grande escala.

Irá obter um único pipeline para copiar rapidamente objetos do armazenamento da origem de dados para o destino de uma forma muito intuitiva.

Cadência de tarefas ou agendamento de tarefas *

☒ Executar uma vez agora ☐ Agendar ☐ Janela em cascata

4.1.1. Criar Ligação

Nova ligação
Azure Data Lake Storage Gen2 [Saiba mais](#)

Nome *

AzureDataLakeStorage1

Descrição

Arquivo de dados de origem

Especifique o arquivo de dados de origem para a tarefa de cópia. Pode utilizar uma ligação de arquivo de dados existente ou especificar um novo.

Tipo de origem Azure Data Lake Storage Gen2

Ligação * AzureDataLakeStorage1 [Editar](#) [Nova ligação](#)

Runtime de integração * ☒ AutoResolveIntegrationRuntime (Recomendado) [Editar](#)

☒ Criação interativa ativada

Ficheiro ou pasta

Se a identidade que utiliza para aceder ao arquivo de dados apenas tem permissão para o subdiretório em vez de toda a conta, especifique o caminho para procurar.

datamvplake/Origem/caso.csv [Procurar](#)

Opções

☐ Cópia binária

☐ Recursivamente

☐ Ativar deteção de partições

4.1.2. Dados de origem

Definições de formato de ficheiro

Formato do ficheiro
DelimitedText

Delimitador de coluna
Comma (,)
Editar

Delimitador de linha
Line feed (\n)
Editar

Avançadas

Tipo de compressão
Selecionar...

Colunas adicionais
+ Novo

Pré-visualizar dados

Serviço associado: AzureDataLakeStorage1

Objeto: datamplake/Origem/caso.csv

Pré-visualização

date	state	city	place_type	confirmed	deaths	order_for_place	is_last	estimated_population_2019	esti
2022-03-27	AP		state	160328	2122	734	True	845731	861
2022-03-26	AP		state	160321	2122	733	False	845731	861
2022-03-25	AP		state	160314	2122	732	False	845731	861
2022-03-24	AP		state	160301	2122	731	False	845731	861
2022-03-23	AP		state	160288	2122	730	False	845731	861
2022-03-22	AP		state	160275	2122	729	False	845731	861

4.1.3. Destino

Arquivo de dados de destino

Especifique o arquivo de dados de destino para a tarefa de cópia. Pode utilizar uma ligação de arquivo de dados existente ou especificar um novo arquivo de dados

Tipo de destino
Azure Synapse Analytics

Ligação *
AzureSynapseAnalytics1 [Editar](#) [+ Nova ligação](#)

Runtime de integração *
☒ AutoResolveIntegrationRuntime (Re... [Editar](#)
☒ Criação interativa ativado ⓘ

Origem
caso

Destino
dbo → caso (criação automática)
[Utilizar a tabela existente](#)

4.1.4. Mapeamento das colunas

Mapeamento de colunas

Escolha a forma de mapeamento das colunas de origem e de destino

Mapeamentos da tabela (1) [+ Novo mapeamento](#) [Limpar](#) [Repor](#) [Eliminar](#)

☒ Origem
Azure Data Lake Storage Gen2 ficheiro
Destino
dbo.caso

	Origem	Tipo		Destino	Tipo	
<input type="checkbox"/>	date	abc String	→	date	abc String	+ -
<input type="checkbox"/>	state	abc String	→	state	abc String	+ -
<input type="checkbox"/>	city	abc String	→	city	abc String	+ -
<input type="checkbox"/>	place_type	abc String	→	place_type	abc String	+ -
<input type="checkbox"/>	confirmed	abc String	→	confirmed	abc String	+ -
<input type="checkbox"/>	deaths	abc String	→	deaths	abc String	+ -
<input type="checkbox"/>	order_for_place	abc String	→	order_for_place	abc String	+ -
<input type="checkbox"/>	is_last	abc String	→	is_last	abc String	+ -
<input type="checkbox"/>	estimated_population...	abc String	→	estimated_population...	abc String	+ -
<input type="checkbox"/>	estimated_population	abc String	→	estimated_population	abc String	+ -
<input type="checkbox"/>	city_ibge_code	abc String	→	city_ibge_code	abc String	+ -
<input type="checkbox"/>	confirmed_per_100k_i...	abc String	→	confirmed_per_100k_i...	abc String	+ -
<input type="checkbox"/>	death_rate	abc String	→	death_rate	abc String	+ -

4.1.5. Pipeline criado



Implementação concluído

Passo de implementação	Estado
A validar o ambiente de runtime da cópia	✓ Efetuado com êxito
> A criar conjuntos de dados	✓ Efetuado com êxito
> A criar pipelines	✓ Efetuado com êxito
> A executar pipelines	✓ Efetuado com êxito

Foram criados conjuntos de dados e pipelines. Agora pode monitorizar e editar os pipelines de cópia ou clicar em terminar para fechar a Ferramenta Copiar Dados.


4.1.6. Executando o Pipeline

All status ▾


A mostrar 1 - 1 de 1 itens

Nome da atividade	Estado de atividade	Tipo de atividade	Início da execução	Duração	Registro
Copy_etu	Em Fila	Copiar dados	9/16/2023, 7:30:56 PM	22s	

ID de execução de atividades: 608c763a-35ce-4cb7-89e6-3d9789d42abb

**Azure Data Lake Storage Gen2**
Região: Brazil South

Efetuated com êxito
Região do IR Azure: Brazil South

**Azure Synapse Analytics**
Região: Brazil South

Dados lidos: 242,108 MB

Ficheiros lidos: 1

Linhas lidas: 2 838 003

Ligações mais elevadas: 11

Dados escritos: 407,542 MB

Linhas escritas: 2 838 003

Ligações mais elevadas: 2

Duração da cópia: 00:11:11
Débito: 388,616 KB/s

▼ Azure Data Lake Storage Gen2 → Azure Synapse Analytics

Hora de início: 2023-09-16T19:30:57.2965158Z
DIUs utilizados: 4
Cópias paralelas utilizadas: 1

▼ Duração: 00:11:11

Detalhes	Duração do trabalho	Duração total
● Fila		00:00:46
● Pré-copiar script		00:00:00
● Transferência		00:10:23
A listar origem	00:00:00	
Leitura a partir da origem	00:00:02	
Escrever no sink	00:10:15	

4.1.7. Tabela DBO.CASO

A tabela “dbo.caso” foi criada com os dados brutos, conforme mencionado no item 4.1, na imagem abaixo podemos verificar que foram inseridas 2.838.003 linhas de dados.

Sinopse em tempo real ▾ | Visualizar todos | Publicar todos

Data | Workspace | Linked

Filtrar recursos por nome

- Base de dados SQL 1
 - PoolMvpSprint2 (SQL)
 - Tabelas
 - dbo.caso
 - Colunas
 - date (nvarchar(max), nu...
 - state (nvarchar(max), n...
 - city (nvarchar(max), null)
 - place_type (nvarchar(m...
 - confirmed (nvarchar(m...
 - deaths (nvarchar(max), ...
 - order_for_place (nvarchar...
 - is_last (nvarchar(max), ...
 - estimated_population_...
 - estimated_population (...
 - city_ibge_code (nvarchar...
 - confirmed_per_100k_in...
 - death_rate (nvarchar(m...

SQL script 1

Executar | Anular | Publicar | Plano de consulta

```
1 SELECT count (*)
2 FROM [dbo].[caso]
```

Resultados | Mensagens

Ver: Tabela | Gráfico | Exportar resultados

Pesquisar

(Sem nome de coluna)

2838003

Resultados

Mensagens

Ver

Tabela

Gráfico

Exportar resultados

Pesquisar

date	state	city	place_type	confirmed	deaths	order_for_place	is_last	estimated_pop...	estimated_pop...	city_ibge_code	confirmed_per...	death_rate
2021-10-05	AP	Laranjal do Jari	city	8398	95	528	False	50410	51362	1600279	16350.60940	0.0113
2021-07-09	AP	Laranjal do Jari	city	8361	95	440	False	50410	51362	1600279	16278.57171	0.0114
2021-04-12	AP	Laranjal do Jari	city	7988	82	352	False	50410	51362	1600279	15552.35388	0.0103
2021-01-14	AP	Laranjal do Jari	city	5433	51	264	False	50410	51362	1600279	10577.85912	0.0094
2020-10-06	AP	Laranjal do Jari	city	4715	45	175	False	50410	51362	1600279	9179.93848	0.0095
2020-07-09	AP	Laranjal do Jari	city	3276	41	86	False	50410	51362	1600279	6378.25630	0.0125
2021-12-10	AP	Macapá	city	61750	1500	619	False	503327	512902	1600303	12039.33695	0.0243
2021-09-09	AP	Macapá	city	60665	1482	528	False	503327	512902	1600303	11827.79556	0.0244
2021-06-10	AP	Macapá	city	54585	1325	437	False	503327	512902	1600303	10642.38393	0.0243
2021-03-11	AP	Macapá	city	37720	870	346	False	503327	512902	1600303	7354.23141	0.0231

5. Tratamento dos dados

Nesta etapa pretendo tratar os dados da tabela `dbo.caso`, modelando os dados a fim de permitir as análises necessárias para responder as perguntas do meu objetivo.

O resultado destes tratamentos será inserido em uma tabela flat chamada “casos” no esquema “covid”.

Antes de seguir fiz algumas pré-análises nos dados brutos e percebi que será possível criar um modelo mais simples, minimizando a quantidade de linhas necessárias para a solução do problema proposto.

Em uma análise inicial identifiquei o seguinte:

- Na coluna **date** verifiquei que a última data de atualização inserida foi dia 27/03/2022.
- A coluna **state** não tem valores nulo, contém as unidades federativas.
- A coluna **city** possui valores nulo quando o registro é referente ao estado.
- Na coluna **place_type** não tem valores nulos, pode existir os valores city (para cidades) e state (para estados).
- A coluna **confirmed** tem os casos confirmados de COVID, tem valor mínimo 0.

Executar

Atualizar

Publicar

Plano de Consulta

1

2

3

4

5

6

7

8

9

10

11

```
SELECT
CAST(confirmed as INT)
FROM [dbo].[caso]

group by CAST(confirmed as INT)

order BY CAST(confirmed as INT) ASC
```

Resultados

Mensagens

Ver

Tabela

Gráfico

Exportar resultados

Pesquisar

(Sem nome de coluna)

0

1

2

3

4

5

- f. A coluna **deaths** tem o número de mortes, tem valor mínimo de 0.

Executar Anular Publicar Plano de consi

```
1 SELECT
2 CAST(deaths as INT)
3 FROM [dbo].[caso]
4
5 group by CAST(deaths as INT)
6
7 order BY CAST(deaths as INT) ASC
8
9
10
11
```

Resultados Mensagens

Ver Tabela Gráfico Exportar resultados

Pesquisar

(Sem nome de coluna)

0
1
2
3
4

- g. A coluna **order_for_place** tem valor mínimo 1, com incremento de 1 a cada boletim.

Executar Anular Publicar Plano de consi

```
1 SELECT
2 CAST(order_for_place as INT)
3 FROM [dbo].[caso]
4
5 group by CAST(order_for_place as INT)
6
7 order BY CAST(order_for_place as INT) ASC
8
9
10
11
```

Resultados Mensagens

Ver Tabela Gráfico Exportar resultados

Pesquisar

(Sem nome de coluna)

1
2
3
4
5
6

- h. A coluna **is_last** não tem valores nulos, contém false ou true. Aqui indica se é ou não o mais novo boletim do local.

city	is_last	date
Abaeté	False	2021-09-20
Abaeté	True	2021-09-22

- i. A coluna **estimated_population_2019** pode ter valor nulo.

```
1 SELECT
2 cast(estimated_population_2019 as int)
3 FROM [dbo].[caso]
4
5 group by cast(estimated_population_2019 as int)
6
7 order by cast(estimated_population_2019 as int) ASC
8
9
10
11
```

Resultados Mensagens

Ver Tabela Gráfico Exportar resultados

Pesquisar

(Sem nome de coluna)

(NULL)

781
837
935
1034

- j. A coluna **estimated_population** pode ter valor nulo.

SQL script 4

Executar Anular Publicar Plano de consulta

```
1 SELECT
2 cast(estimated_population as int)
3 FROM [dbo].[casos]
4
5 group by cast(estimated_population as int)
6
7 order by cast(estimated_population as int) ASC
8
9
10
11
```

Resultados Mensagens

Ver Tabela Gráfico Exportar resultados

Pesquisar

(Sem nome de coluna)
(NULL)
776
838
946
982
1118
1153

- k. A coluna **city_ibge_code** pode ter valor nulo nos casos em que “city” tiver valor como Importados/Indefinidos.

Resultados Mensagens

Ver Tabela Gráfico Exportar resultados

Pesquisar

city	(Sem nome de coluna)
Importados/Indefinidos	(NULL)
(NULL)	11
(NULL)	12
(NULL)	13
(NULL)	14

- l. A coluna **confirmed_per_100k_inhabitants** pode ter valor nulo.

Executar Anular Publicar Plano de consulta Ligar a PoolMvpSprint2

```
1 SELECT top 100
2 city,
3 cast(confirmed_per_100k_inhabitants as FLOAT) confirmed_per_100k_inhabitants
4 FROM [dbo].[casos]
5
6 group by city, cast(confirmed_per_100k_inhabitants as FLOAT)
7
8 order by cast(confirmed_per_100k_inhabitants as FLOAT) ASC
9
10
11
```

Resultados Mensagens

Ver Tabela Gráfico Exportar resultados

Pesquisar

city	confirmed_per_100k_inhabitants
Bom Jesus da Penha	(NULL)
Bom Jesus da Lapa	(NULL)
Bom Jesus	(NULL)
Bom Jardim de Goiás	(NULL)
Bodoquena	(NULL)

m. A coluna **death_rate** tem valor mínimo 0.

SQL script 4

Executar Anular Publicar Plano de consulta

```

1 SELECT top 100
2
3 cast(death_rate as FLOAT) death_rate
4 FROM [dbo].[caso]
5
6 group by cast(death_rate as FLOAT)
7
8 order by cast(death_rate as FLOAT) ASC
9
10
11

```

Resultados Mensagens

Ver Tabela Gráfico Exportar resultados

Pesquisar

death_rate

0

0.0001

0.0003

0.0004

0.0005

Etapas que seguirei após as análises feitas dos dados:

- utilizar os últimos dados atualizados, já que a coluna **is_last** nos permite filtrar os registros mais recentes.
- filtrar apenas os estados, desconsiderando as linhas com dados por municípios utilizando um filtro na coluna **place_type**.
- desconsiderar as colunas **city**, **place_type**, **estimated_population_2019**, **is_last**, **order_for_place**, **death_rate**.

5.1. ETL



- Origem: importando os dados de `dbo.caso`, aplicando uma SQL para filtrar apenas os estados e as últimas atualizações.
- Seleção: Manter apenas as colunas `date`, `state`, `confirmed`, `deaths`, `estimated_population`, `city_ibge_code` e `confirmed_per_100k_inhabitants`.
- Converter: mudança no conjunto de valores.

Antes

Ordenar ↑↓	Coluna ↑↓	Tipo ↑↓
1	date	abc string
2	state	abc string
3	confirmed	abc string
4	deaths	abc string
5	estimated_population	abc string
6	city_ibge_code	abc string
7	confirmed_per_100k_inhabitants	abc string

Depois

Número de colunas Novo 0		Inalterado 1
Ordenar ↑↓	Coluna ↑↓	Tipo ↑↓
1	date	date
2	state	string
3	confirmed	long
4	deaths	long
5	estimated_population	long
6	city_ibge_code	long
7	confirmed_per_100k_inhabitants	double

PoolMvpSprint2 (SQL)

Tabelas

covid.casos

Colunas

- date (date, null)
- state (nvarchar(max), null) ...
- confirmed (bigint, null)
- deaths (bigint, null)
- estimated_population (bigint, null)
- city_ibge_code (bigint, null)
- confirmed_per_100k_inhabitants (float, null)

Como resultado reduzi de mais de 2 milhões de linhas para apenas 27, o que corresponde a quantidade de unidades federativas existentes.

Executar Anular | Publicar

```
1 SELECT count(*)
2 FROM covid.casos
```

Resultados Mensagens

Ver Tabela Gráfico Exportar

Pesquisar


(Sem nome de coluna)

27


6. Data Catalog


Utilizei a ferramenta **Microsoft PurView** para catalogar os dados.


Collection path


 [MVPPurView](#)

Hierarchy

 [mvplake](#)
Azure Synapse Workspace

 [PoolMvpSprint2](#)
Azure Synapse Dedicated SQL Database

 [covid](#)
Azure Synapse Dedicated SQL Schema

 [casos](#)
Azure Synapse Dedicated SQL Table


Properties

name	casos
principalId	0
qualifiedName	mssql://mvplake.sql.azuresynapse.net/PoolMvpSprint2/covid/casos

Related assets

DbSchema	covid
----------	-----------------------

Data catalog >

 **casos**
Azure Synapse Dedicated SQL Table
[+ Add Tag](#)

[Edit](#) [Select for bulk edit](#) [Request access](#) [Refresh](#) [Delete](#) [Edit columns](#)

[Overview](#) [Properties](#) [Schema](#) [Lineage](#) [Contacts](#) [Related](#)

Showing 7 of 7 items

Column name	Classifications	Data type	Column description
date		date	data de coleta dos dados no formato YYYY-MM-DD
state		nvarchar	sigla da unidade federativa, exemplo: SP
confirmed		bigint	número de casos de COVID-19 confirmados
deaths		bigint	número de mortes por COVID-19
estimated_population		bigint	população estimada para esse município/estado em 2020
city_ibge_code		bigint	código IBGE do local
confirmed_per_100k_inhabitants		float	número de casos confirmados por 100.000 habitantes (baseado em estimated_population)

7. Conclusão

No início deste projeto tentei utilizar a AWS da Amazon, mas durante o processo encontrei dificuldades, a plataforma começou a cobrar para utilizá-la e, como segunda alternativa, passei a utilizar o Azure da Microsoft, o que na minha opinião se melhor encaixou na resolução do meu problema de modelagem.

Para o processo de ingestão, ETL e análise optei por utilizar a Azure Synapse Analytics, um serviço com intuito de acelerar o tempo de insight entre Data Warehouse e sistemas de Big Data.

No processo de ETL foi possível a redução de mais de 2 milhões de linhas da base original para 27, uma linha para cada unidade federativa, além de reduzir de 13 para 7 atributos.

Esta estratégia foi utilizada para melhor se encaixar o modelo na solução do problema, já que a análise seria por cada estado brasileiro.

Durante a pré-análise dos dados encontrei algumas colunas que aceitavam dados nulos, mas isso não foi problema para o nosso objetivo principal, já que os atributos que seriam utilizados tinham dados inseridos e íntegros. De forma geral o conjunto de dados estava tratado, o que agilizou todo o processo de modelagem.

Agora chegou o momento tão esperado, está na hora de solucionar o nosso problema. Nesta etapa vou efetuar uma análise dos dados da tabela "casos" para responder as perguntas do nosso objetivo.

Começamos dando um resumo das características da tabela "casos" do esquema "covid" criada:

- Tabela do tipo flat.
- Contém 7 atributos.
- 27 linhas.
- Sem valores nulos.
- 1 coluna não numérica e 6 colunas numéricas.
- Última atualização feita no dia 27/03/2022.

date	state	confirmed	deaths	estimated_population	city_ibge_code	confirmed_per_100k_inhabi...
2022-03-26T00:00:00.0000000	RN	495749	8119	3534165	24	14027.33036
2022-03-27T00:00:00.0000000	RS	2263880	38985	11422973	43	19818.65842
2022-03-27T00:00:00.0000000	MT	724653	14854	3526220	51	20550.41943
2022-03-27T00:00:00.0000000	AP	160328	2122	861773	16	18604.43527
2022-03-26T00:00:00.0000000	RR	155062	2144	631181	14	24566.96257
2022-03-27T00:00:00.0000000	SP	5232374	167110	46289333	35	11303.62799
2022-03-27T00:00:00.0000000	PA	751293	18079	8690745	15	8644.74795
2022-03-27T00:00:00.0000000	AC	123808	1992	894470	12	13841.49273
2022-03-27T00:00:00.0000000	RO	391943	7172	1796460	11	21817.5189

Para responder as perguntas dos objetivos optei por utilizar consultas SQL, nesta etapa não encontrei dificuldades.

Meu objetivo proposto foram 7 perguntas relacionadas à pandemia da COVID, um momento muito delicado que a humanidade passou e ainda passa, mas graças à medicina moderna estamos conseguindo contornar.

Basicamente a ideia é entender os casos de contágio e mortes no cenário nacional, e a seguir as respostas que encontrei:

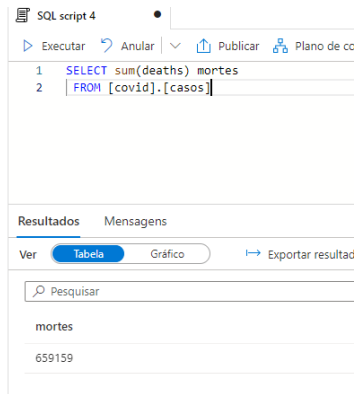
a. Quantos casos de Covid-19 confirmados?

Temos 29.849.740 casos confirmados em todos os estados.

confirmados
29849740

b. Quantas mortes por causa da Covid?

São 659.159 mortes confirmadas.

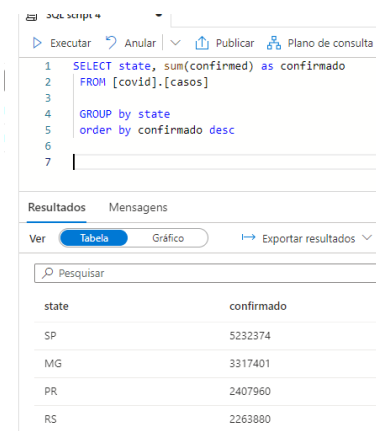


The screenshot shows a SQL script editor with a query to count the number of deaths. The query is: `SELECT sum(deaths) mortes FROM [covid].[casos]`. The results tab shows a single row with the value 659159.

Ver	Tabela	Gráfico	Exportar resultado
Pesquisar			
mortes			
659159			

c. Qual é o estado com maior volume de casos de Covid confirmados?

O estado de São Paulo possui o maior volume de casos confirmados.

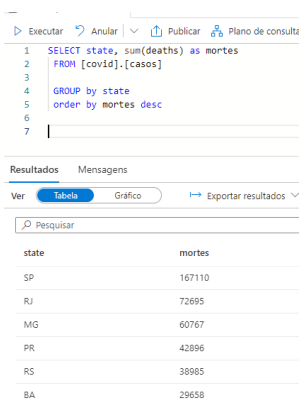


The screenshot shows a SQL script editor with a query to find the state with the most confirmed cases. The query is: `SELECT state, sum(confirmed) as confirmado FROM [covid].[casos] GROUP by state order by confirmado desc`. The results tab shows a table with two columns: state and confirmado. The data is as follows:

state	confirmado
SP	5232374
MG	3317401
PR	2407960
RS	2263880

d. Qual é o estado com maior número de mortes?

O estado de São Paulo possui o maior número de mortes.



The screenshot shows a SQL script editor with a query to find the state with the most deaths. The query is: `SELECT state, sum(deaths) as mortes FROM [covid].[casos] GROUP by state order by mortes desc`. The results tab shows a table with two columns: state and mortes. The data is as follows:

state	mortes
SP	167110
RJ	72695
MG	60767
PR	42896
RS	38985
BA	29658

e. Qual estado com o maior índice de morte em relação aos casos confirmados?

O estado do Rio de Janeiro se apresentou como o de maior número de mortos em relação ao número de casos de COVID confirmados.

1	SELECT state, sum(deaths) as mortes,
2	sum(confirmed) as confirmados,
3	cast(sum(deaths) as FLOAT)/cast(sum(confirmed) as FLOAT) *100 as death_rate
4	FROM [covid].[casos]
5	
6	GROUP by state
7	order by death_rate desc
8	

state	mortes	confirmados	death_rate
RJ	72695	2078817	3.49694080816157
SP	167110	5232374	3.1937701700987
MA	10869	424199	2.56224063507976
AM	14151	581070	2.425334811336156
PA	18079	751293	2.4063847260656
PE	21366	892115	2.39496270962824
AL	8800	368073	2.39077345450316

f. Qual estado com o menor índice de morte em relação aos casos confirmados?

Neste item o estado de Santa Catarina ficou em destaque.

1	SELECT state, sum(deaths) as mortes,
2	sum(confirmed) as confirmados,
3	cast(sum(deaths) as FLOAT)/cast(sum(confirmed) as FLOAT) *100 as death_rate
4	FROM [covid].[casos]
5	
6	GROUP by state
7	order by death_rate asc
8	

state	mortes	confirmados	death_rate
SC	21648	1671175	1.29537600789863
AP	2122	160328	1.32353674966319
TO	4142	302502	1.36924714547342
ES	14323	1037188	1.38094540237643
RR	2144	155062	1.38267273735667
AC	1992	123808	1.60894287929698
DF	8110	405240	1.97773207008307

g. Qual estado com o menor índice de contágio para cada 100 mil habitantes?

O estado do Maranhão se apresentou com o menor índice de casos a cada 100 mil habitantes.

4	cast(sum(confirmed) as FLOAT)/cast(sum(estimated_population) as FLOAT) *100000 as contagio_por_100mil_habi
5	confirmed_per_100k_inhabitants
6	FROM [covid].[casos]
7	
8	GROUP by state, confirmed_per_100k_inhabitants
9	order by contagio_por_100mil_habitante ASC
10	
11	

state	confirmados	nro_habitantes	contagio_por_100mil_habitante
MA	424199	7114598	5962.37482426976
PA	751293	8690745	8644.74794738541
AL	295972	3351543	8830.9175803503
PE	892115	9616621	9276.80315154356
BA	1530054	14930634	10247.7496936835
PI	367515	3281480	11199.6720991748
SP	5232374	46289333	11303.6279870354
RJ	2078817	17366189	11970.4847160192