

# MVP – Engenharia de Dados

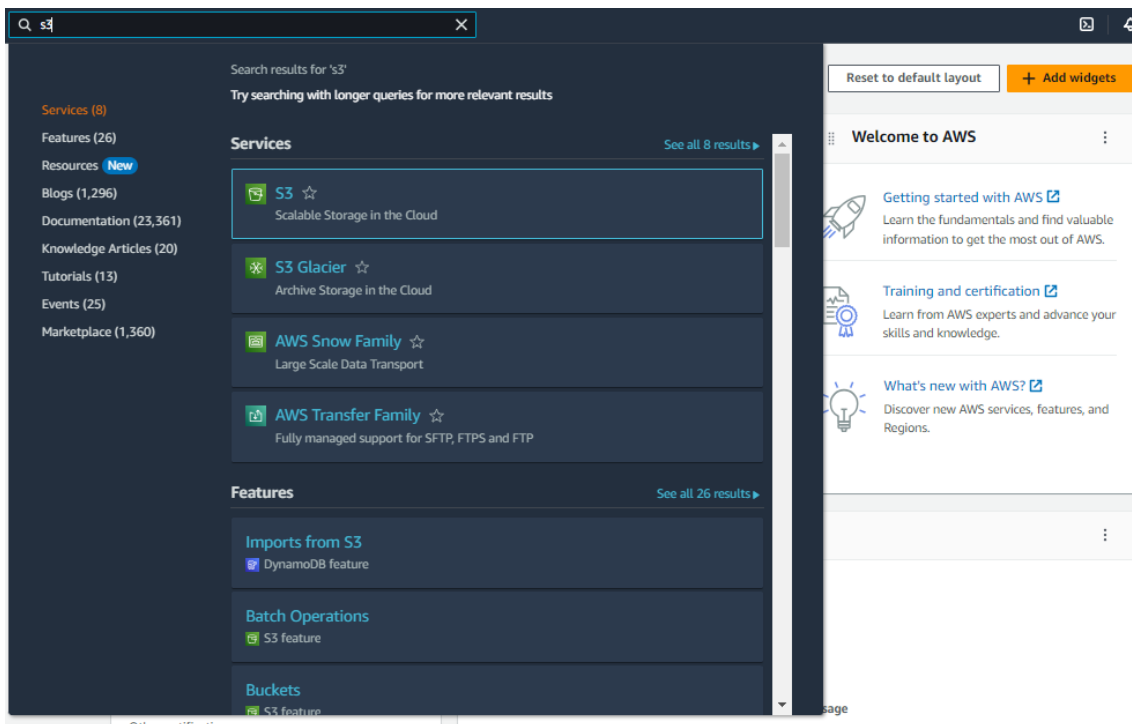
Josafá de Sousa Gomes Júnior

## 1. Objetivo

Com os dados obtidos da nuvem e uso de SQL, iremos retirar alguns insights que serão mostrados diretamente do banco contendo dados da Covid 19 no Brasil retirados do website <https://brasil.io/dataset/covid19/files/>. O arquivo que foi feito o tratamento está no formato CSV.

## 2. Modelagem

AWS Glue e Amazon Athena serão usados para fazer a leitura dos dados e realizar queries para extrair informações desejadas dos bancos de forma a extrair o que pode ser agregador na leitura e entendimento do mesmo.



Acessar o S3 e criar um bucket

cnpathena [Info](#)

[Objects](#) | [Properties](#) | [Permissions](#) | [Metrics](#) | [Management](#) | [Access Points](#)

---

**Objects (1)**

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

[Refresh](#) [Copy S3 URI](#) [Copy URL](#) [Download](#) [Open](#) [Delete](#) [Actions](#)

[Create folder](#) [Upload](#)

< 1 > [Settings](#)

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	<a href="#">covid-brasil/</a>	Folder	-	-	-

Nome dado ao bucket: cnpathena

[Amazon S3](#) > [Buckets](#) > cnpathena

cnpathena [Info](#)

[Objects](#) | [Properties](#) | [Permissions](#) | [Metrics](#) | [Management](#) | [Access Points](#)

---

**Objects (1)**

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

[Refresh](#) [Copy S3 URI](#) [Copy URL](#) [Download](#) [Open](#) [Delete](#) [Actions](#)

[Create folder](#) [Upload](#)

< 1 > [Settings](#)

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	<a href="#">covid-brasil/</a>	Folder	-	-	-

Criada uma pasta dentro do bucket chamada: covid-brasil

covid-brasil/ [Copy S3 URI](#)

**Objects** | Properties

**Objects (1)**  
Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

[Refresh](#) [Copy S3 URI](#) [Copy URL](#) [Download](#) [Open](#) [Delete](#) [Actions](#) [Create folder](#)

[Upload](#)

< 1 > ⚙

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	caso_full.csv	csv	September 27, 2023, 21:10:39 (UTC-03:00)	417.5 MB	Standard

Feito o upload do arquivo csv na pasta covid-brasil (tamanho: 417.5 MB)

cnpathenaresults [Info](#)

**Objects** | Properties | Permissions | Metrics | Management | Access Points

**Objects (0)**  
Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

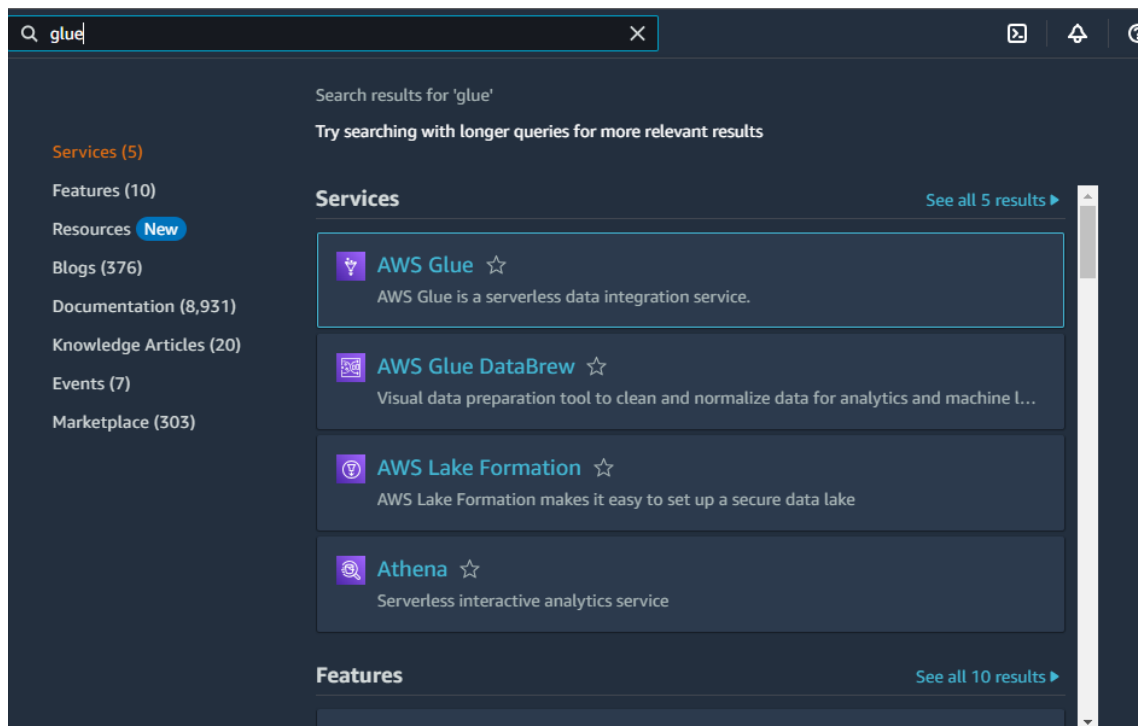
[Refresh](#) [Copy S3 URI](#) [Copy URL](#) [Download](#) [Open](#) [Delete](#) [Actions](#)

[Create folder](#) [Upload](#)

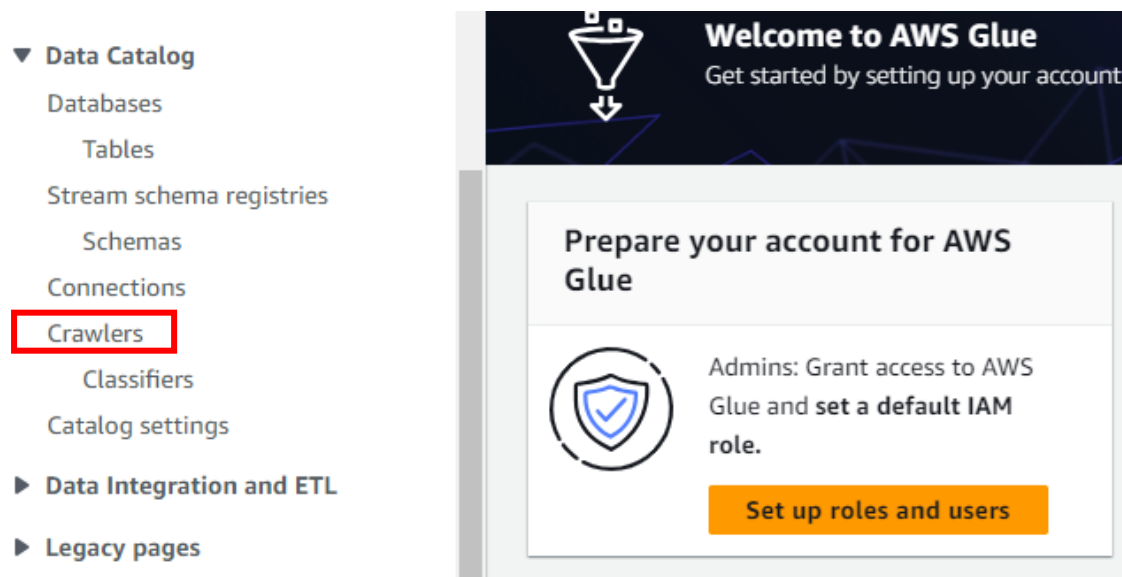
< 1 > ⚙

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
--------------------------	------	------	---------------	------	---------------

Novo bucket criado e adicionado para armazenar os resultados da query



Acessar o AWS Glue para fazer a parte de ETL (Tratamento de dados) com o uso de Crawlers



Inspeciona os arquivos do bucket e pega os arquivos csv para criar uma tabela

## Crawlers

A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

Crawlers (0) [Info](#)

View and manage all available crawlers.

Last updated (UTC)  
September 28, 2023 at 24:32:07



Action ▼

Run

Create crawler

Filter crawlers

< 1 > ⚙

	Name	State	Schedule	Last run	Last run ti...	Log	Table change...
--	------	-------	----------	----------	----------------	-----	-----------------

No resources

No resources to display.

Início do processo de configuração para criar o crawler

## Configure security settings

IAM role [Info](#)

Existing IAM role

AWSGlueServiceRole-athena



View [↗](#)

Create new IAM role

Update chosen IAM role

Only IAM roles created by the AWS Glue console and have the prefix "AWSGlueServiceRole-" can be updated.

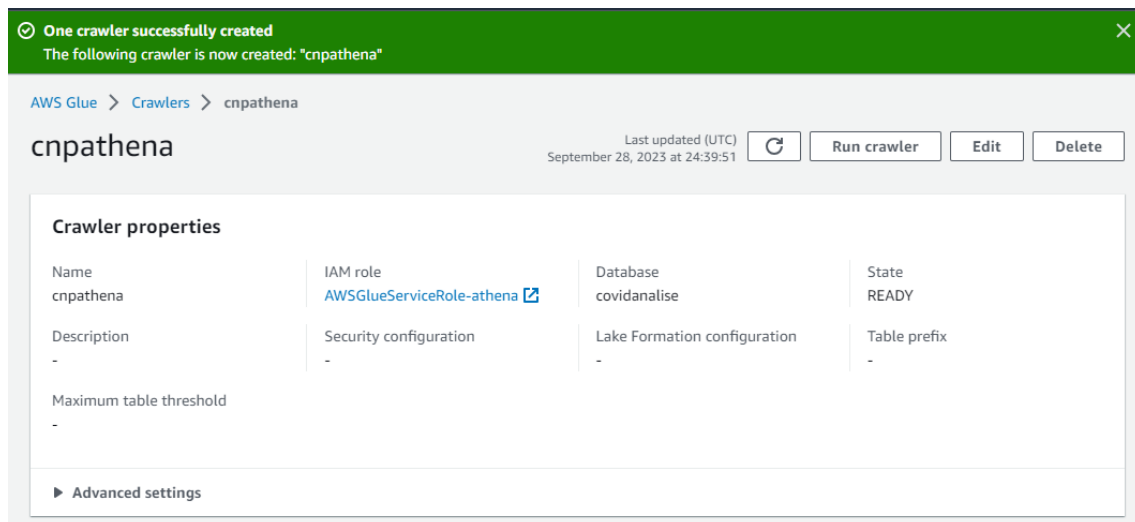
### Lake Formation configuration - *optional*

Allow the crawler to use Lake Formation credentials for crawling the data source. [Learn more.](#) [↗](#)

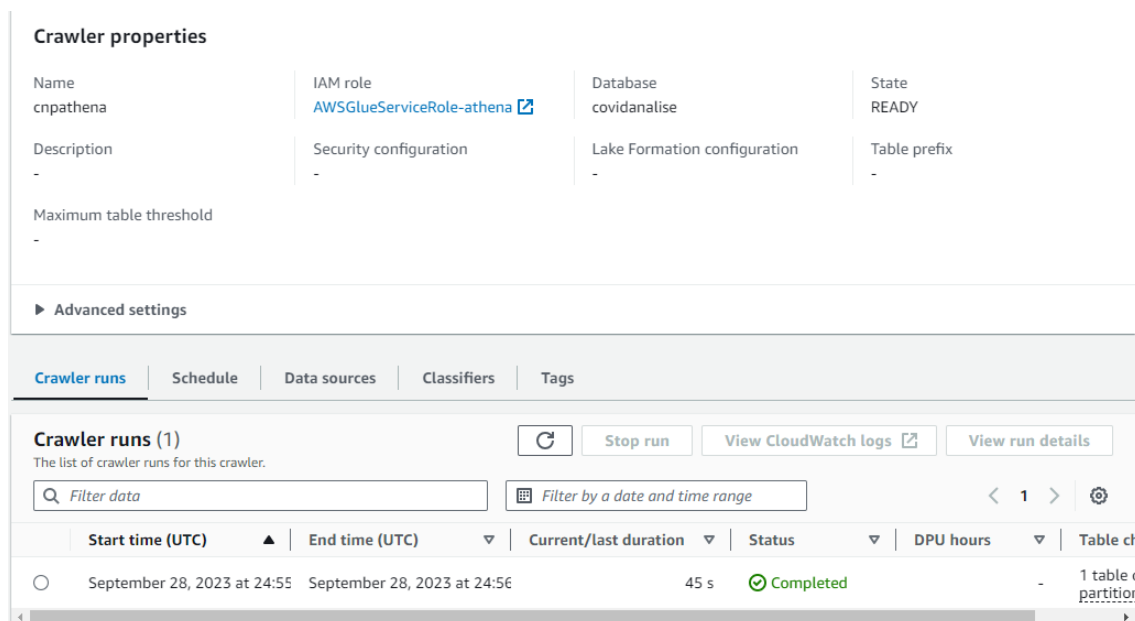
☐ Use Lake Formation credentials for crawling S3 data source

Checking this box will allow the crawler to use Lake Formation credentials for crawling the data source. If the data source is registered in another account, you must provide the registered account ID. Otherwise, the crawler will crawl only those data sources associated to the account. Only applicable to S3, Glue Catalog, Iceberg, and Hudi data sources.

Nomear o IAM para habilitar a comunicação entre os serviços, ou seja, com o S3



Tela mostrando que o crawler foi criado com sucesso



Após clicar em "Run Crawler" o arquivo é executado no mesmo

▼ Data Catalog

Databases

**Tables**

Stream schema registries

Schemas

Connections

Crawlers

Classifiers

Catalog settings

► Data Integration and ETL

► Legacy pages

What's New

Documentation

AWS Marketplace

Enable compact mode

cnpathena

[AWSGlueServiceRole-athena](#)

covidanalise

READY

Description

Security configuration

Lake Formation configuration

Table prefix

Maximum table threshold

► Advanced settings

Crawler runs

Schedule

Data sources

Classifiers

Tags

Crawler runs (1)

The list of crawler runs for this crawler.

Filter data

Filter by a date and time range

Start time (UTC)

End time (UTC)

Current/last duration

Status

DPU hours

Table d

September 28, 2023 at 24:55

September 28, 2023 at 24:56

45 s

Completed

1 table

partition

Passo seguinte é clicar em tabelas

covid\_brasil

Table overview

Data quality New

Table details

Advanced properties

Name

covid\_brasil

Location

s3://cnpathena/covid-brasil/

Input format

org.apache.hadoop.mapred.TextInputFormat

Schema

Partitions

Indexes

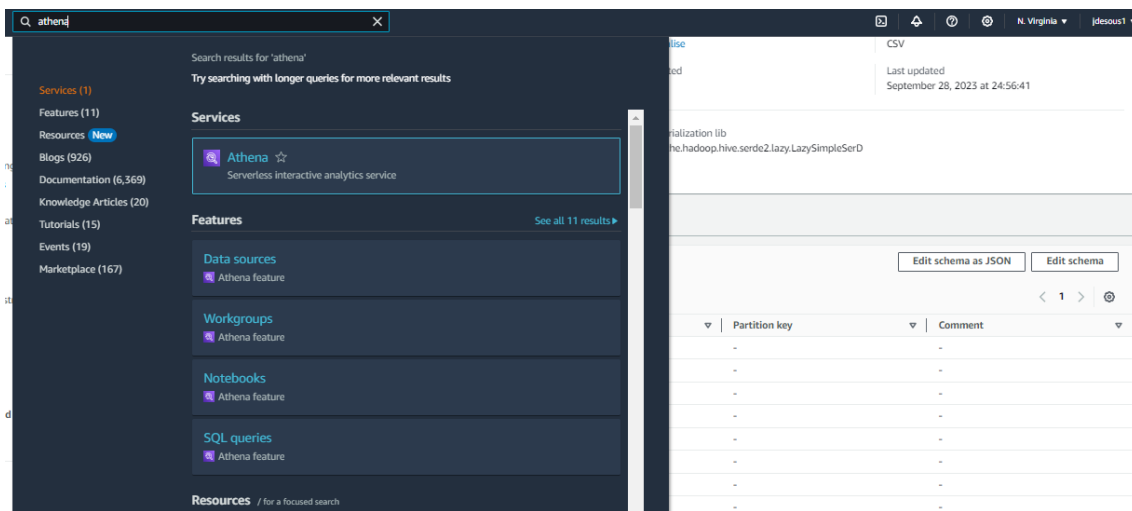
Schema (18)

View and manage the table schema.

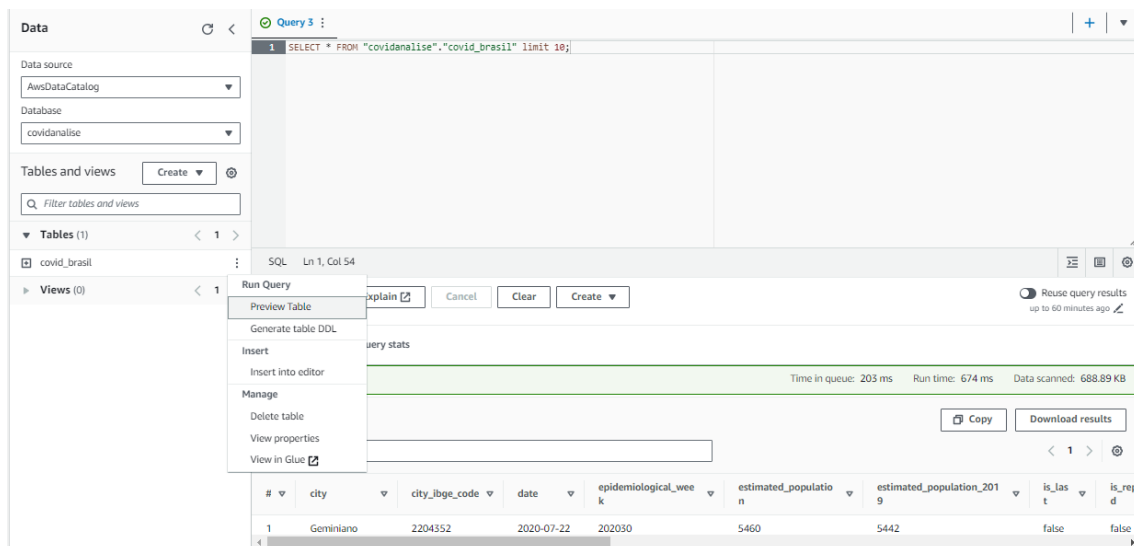
Filter schemas

#	Column name	Data type
1	city	string
2	city_ibge_code	bigint
3	date	string
4	epidemiological_week	bigint
5	estimated_population	bigint
6	estimated_population_2...	bigint
7	is_last	boolean
8	is_repeated	boolean
9	last_available_confirmed	bigint
10	last_available_confirmed...	double
11	last_available_date	string
12	last_available_death_rate	double
13	last_available_deaths	bigint
14	order_for_place	bigint
15	place_type	string
16	state	string
17	new_confirmed	bigint
18	new_deaths	bigint

O esquema mostrado apresenta as colunas presentes no arquivo e os tipos de variáveis contidos nele



Passo seguinte é acessar o Athena para fazer queries no arquivo csv



Antes de rodar a query, configurar onde o Athena irá fazê-lo, dessa forma o Athena ficará configurado e poderá ser exibido do lado esquerdo a tabela covid\_brasil, clicando nos três pontos e em “preview table”, é possível rodar um preview da tabela.



Completed Time in queue: 203 ms Run time: 674 ms Data scanned: 688.89 KB

Results (10) Copy Download results

Search rows

#	city	city_ibge_code	date	epidemiological_week	estimated_population	estimated_population_2019	is_latest	is_reported
1	Geminiano	2204352	2020-07-22	202030	5460	5442	false	false
2	Gilbués	2204402	2020-07-22	202030	10694	10690	false	false
3	Guadalupe	2204501	2020-07-22	202030	10497	10499	false	false
4	Guaribas	2204550	2020-07-22	202030	4568	4562	false	false
5	Hugo Napoleão	2204600	2020-07-22	202030	3879	3877	false	false
6	Ilha Grande	2204659	2020-07-22	202030	9457	9426	false	false
7	Inhuma	2204709	2020-07-22	202030	15319	15308	false	false
8	Ipiranga do Piauí	2204808	2020-07-22	202030	9838	9811	false	false
9	Isaías Coelho	2204907	2020-07-22	202030	8566	8549	false	false
10	Itainópolis	2205003	2020-07-22	202030	11551	11530	false	false

Tabela criada a partir da query executada

### 3. Solução do Problema

A partir da modelagem realizada vamos retirar insights dos dados modelados. Aplicando SQL temos as seguintes informações:

SQL Ln 1, Col 51

Run again Explain Cancel Clear Create Reuse query results up to 60 minutes ago

Query results Query stats

Completed Time in queue: 211 ms Run time: 1.5 sec Data scanned: 417.46 MB

Results (1) Copy Download results

Search rows

#	Total_de_Cidades
1	5597

Total de cidades afetadas pela Covid 19 com base na tabela exploradas

Query 6 : X

Query 7 : X

+ | ▼

1 SELECT estimated\_population FROM "covidanalise"."covid\_brasil"

2 where city = 'São Paulo'

3 limit 1

SQL Ln 3, Col 8

Run again

Explain

Cancel

Clear

Create ▼

Reuse query results

up to 60 minutes ago

Query results

Query stats

Completed

Time in queue: 230 ms

Run time: 873 ms

Data scanned: 288.02 MB

Results (1)

Copy

Download results

Search rows

< 1 > ⚙

# ▼ estimated\_population ▼

1 12325232

Analisando a cidade de São Paulo dentro da amostra, temos que a população estimada até a data da última coleta de dados é de 12.325.232 habitantes

Query 6 : X | Query 7 : X

```

1 SELECT estimated_population_2019 FROM "covidanalise"."covid_brasil"
2 where city = 'São Paulo'
3 limit 1

```

SQL Ln 1, Col 33

Run again Explain Cancel Clear Create

Reuse query results up to 60 minutes ago

Query results | Query stats

Completed Time in queue: 206 ms Run time: 729 ms Data scanned: 142.71 MB

Results (1) Copy Download results

Search rows

#	estimated_population_2019
1	12252023

A população estimada em 2019 era de 12.252.023 habitantes

Query 6 : X | Query 7 : X

```

1 SELECT estimated_population - estimated_population_2019 as Result
2 FROM "covidanalise"."covid_brasil"
3 where city = 'São Paulo'
4 limit 1;

```

SQL Ln 4, Col 9

Run again Explain Cancel Clear Create

Reuse query results up to 60 minutes ago

Query results | Query stats

Completed Time in queue: 122 ms Run time: 617 ms Data scanned: 152.67 MB

Results (1) Copy Download results

Search rows

#	Result
1	73209

Total de mortes confirmadas: 73209

