MVP – Engenharia de Dados

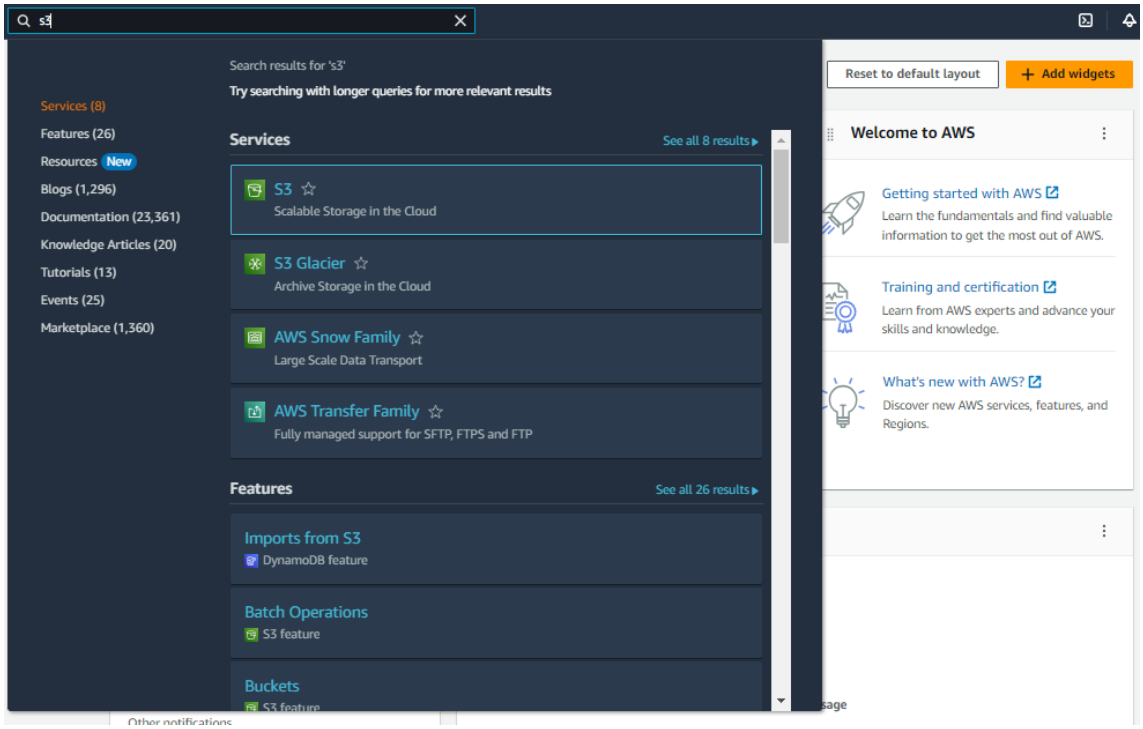
Josafá de Sousa Gomes Júnior

1. Objetivo

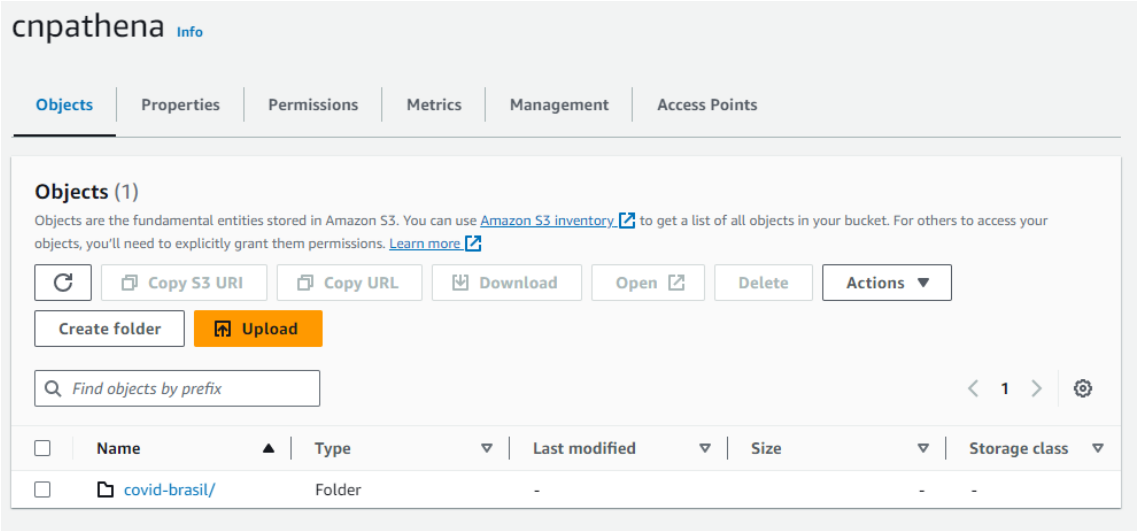
Com os dados obtidos da nuvem e uso de SQL, iremos retirar alguns insights que serão mostrados diretamente do banco contendo dados da Covid 19 no Brasil retirados do website https://brasil.io/dataset/covid19/files/. O arquivo que foi feito o tratamento está no formato CSV.

2. Modelagem

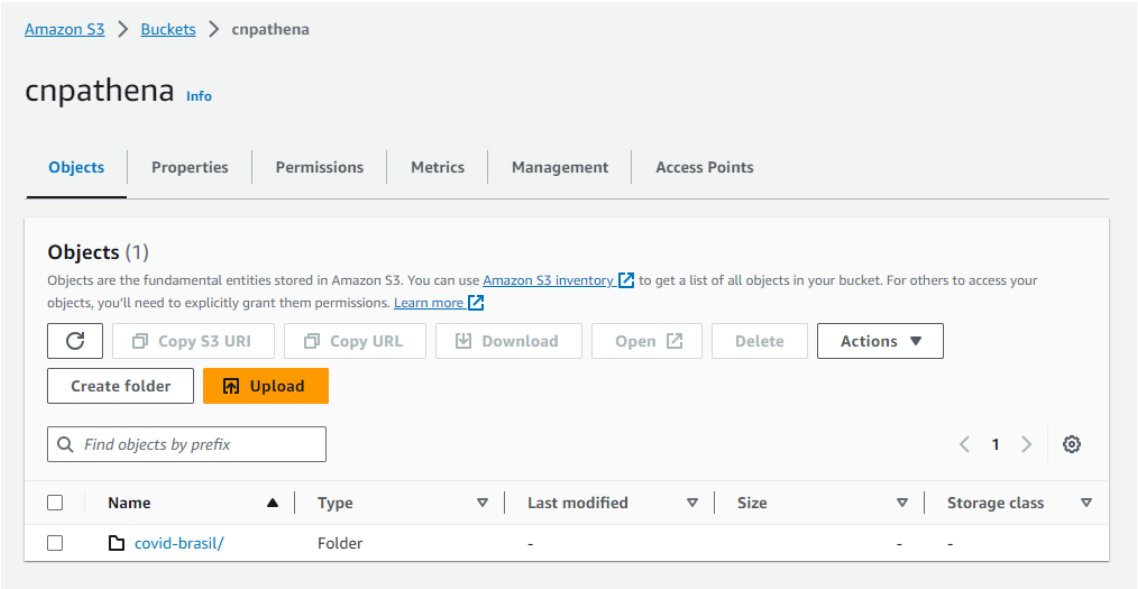
AWS Glue e Amazon Athena serão usados para fazer a leitura dos dados e realizar queries para extrair informações desejadas dos bancos de forma a extrair o que pode ser agregador na leitura e entendimento do mesmo.



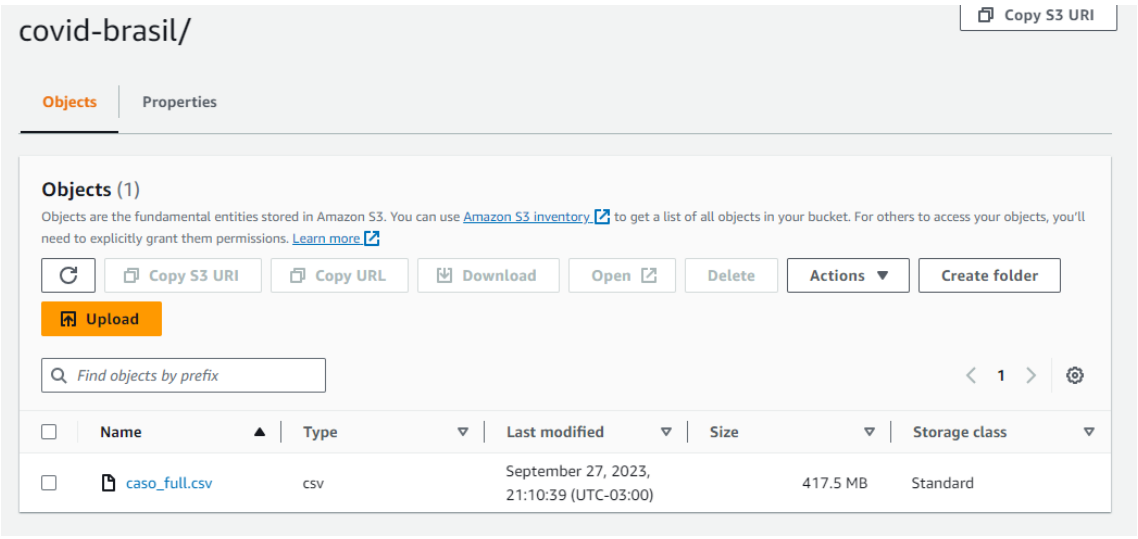
Acessar o S3 e criar um bucket



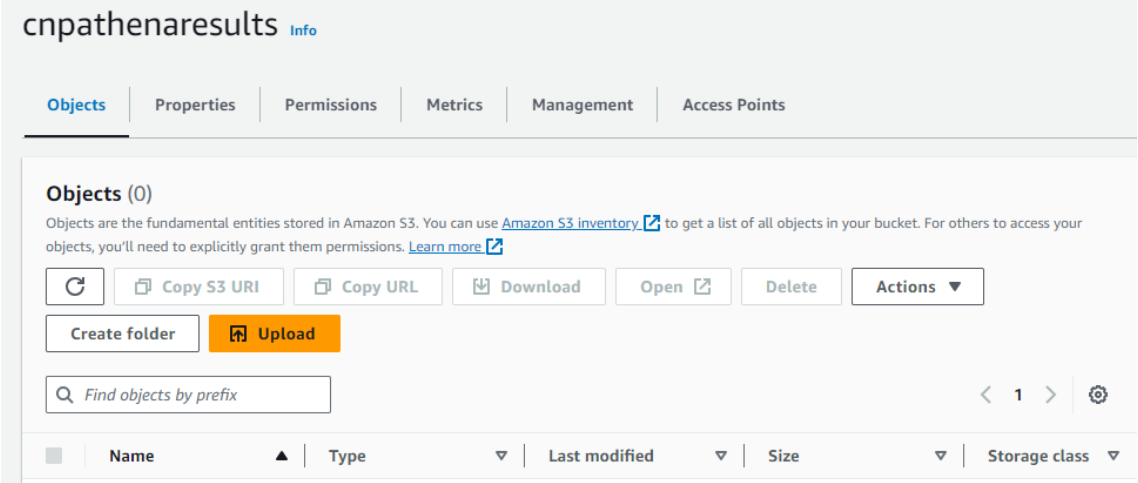
Nome dado ao bucket: cnpathena



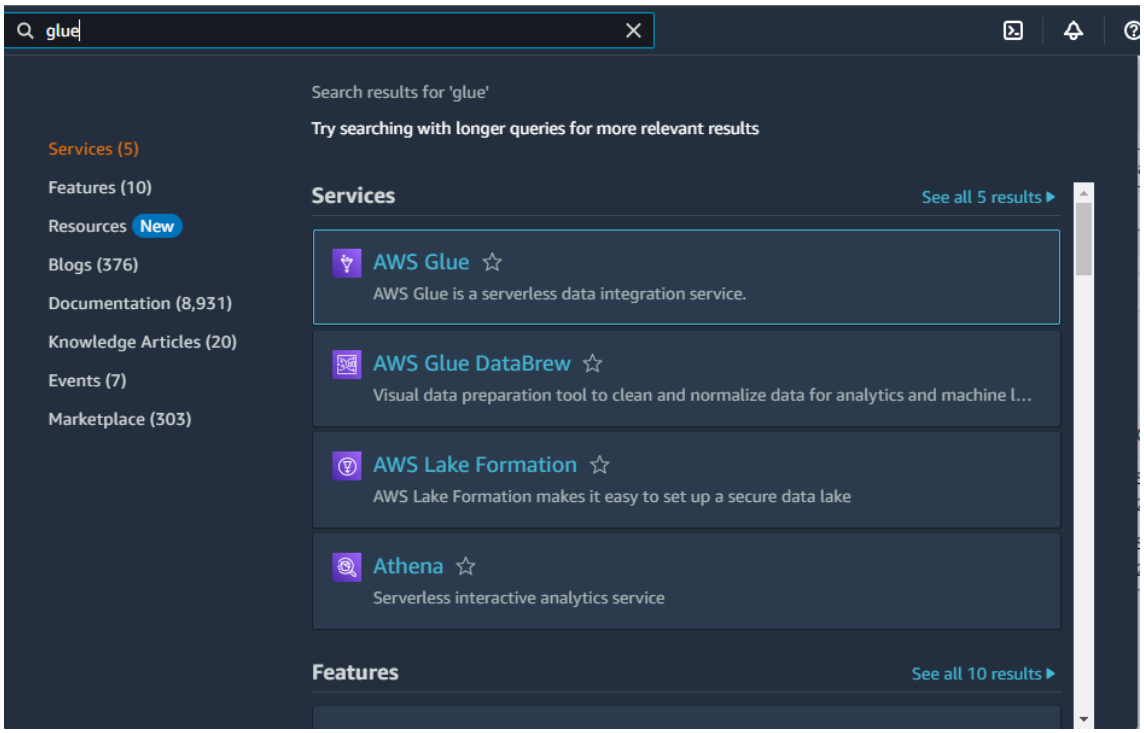
Criada uma pasta dentro do bucket chamada: covid-brasil



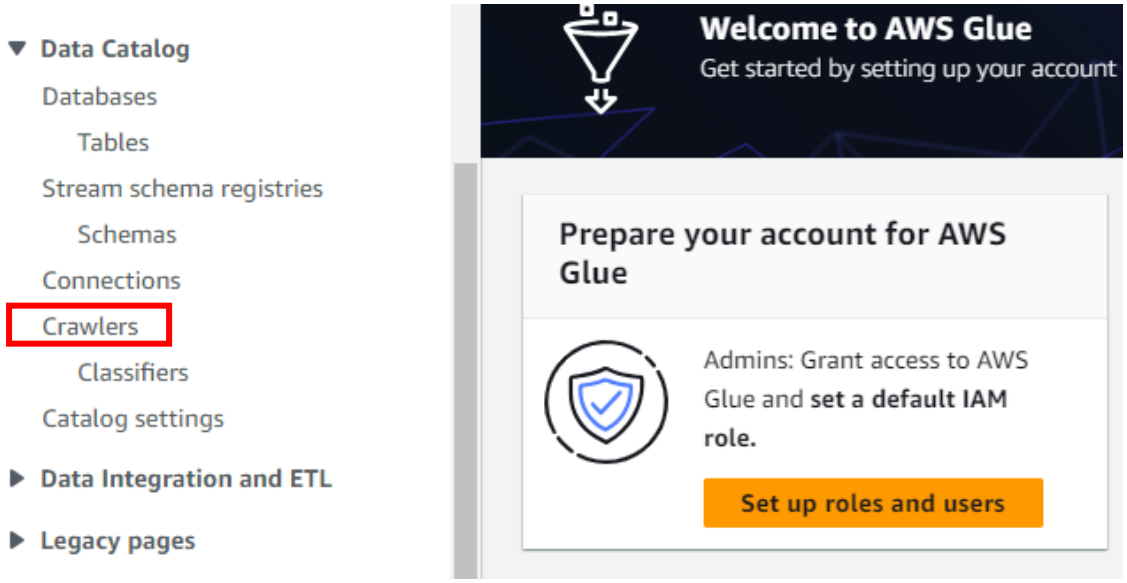
Feito o upload do arquivo csv na pasta covid-brasil (tamanho: 417.5 MB)



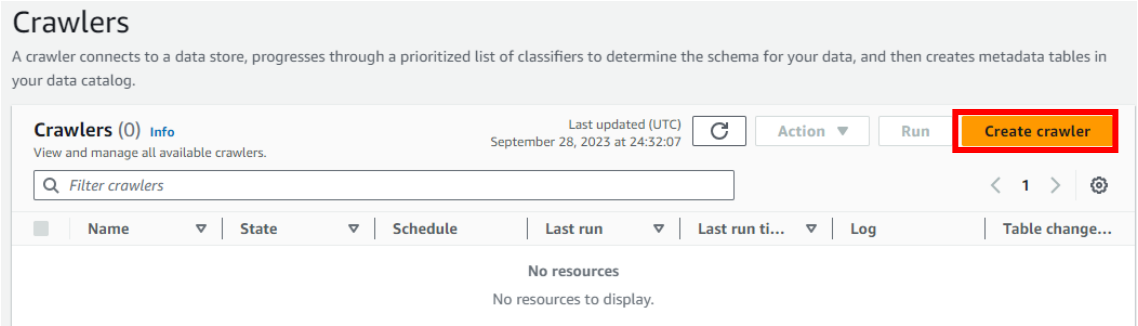
Novo bucket criado e adicionado para armazenar os resultados da query



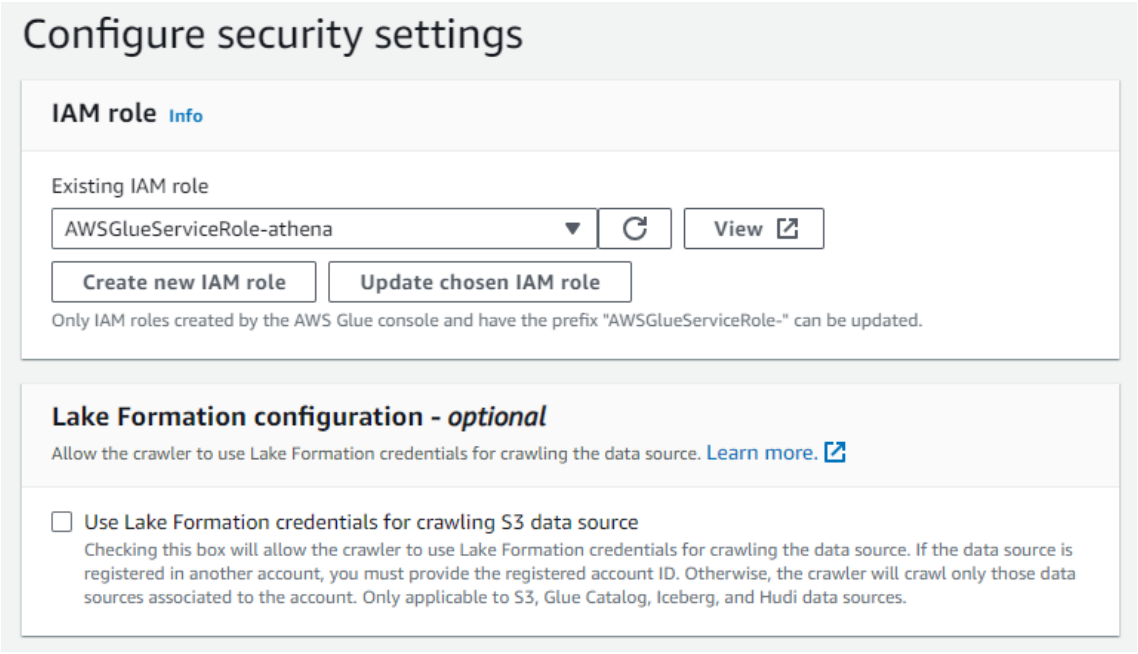
Acessar o AWS Glue para fazer a parte de ETL (Tratamento de dados) com o uso de Crawlers



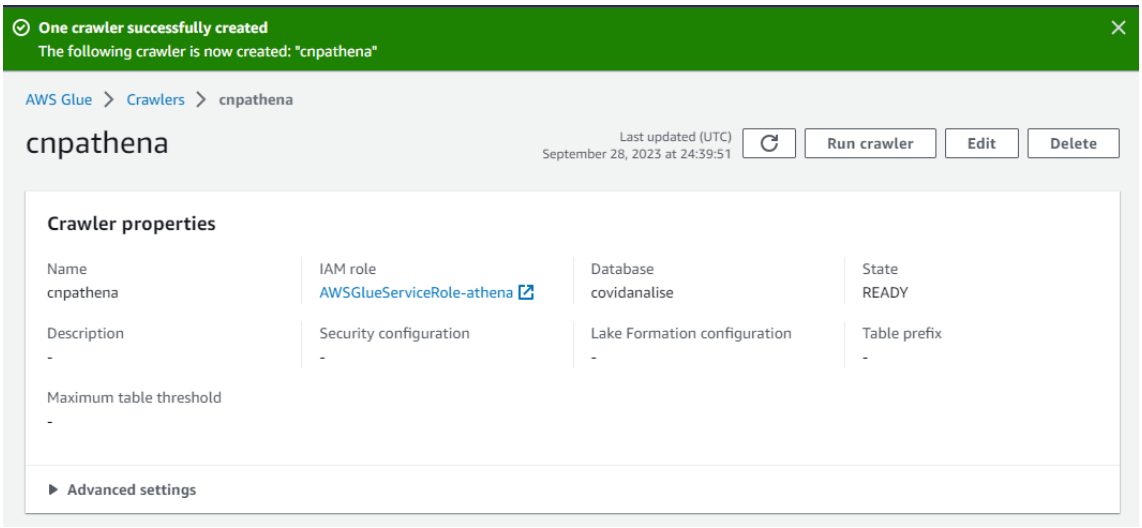
Inspeciona os arquivos do bucket e pega os arquivos csv para criar uma tabela



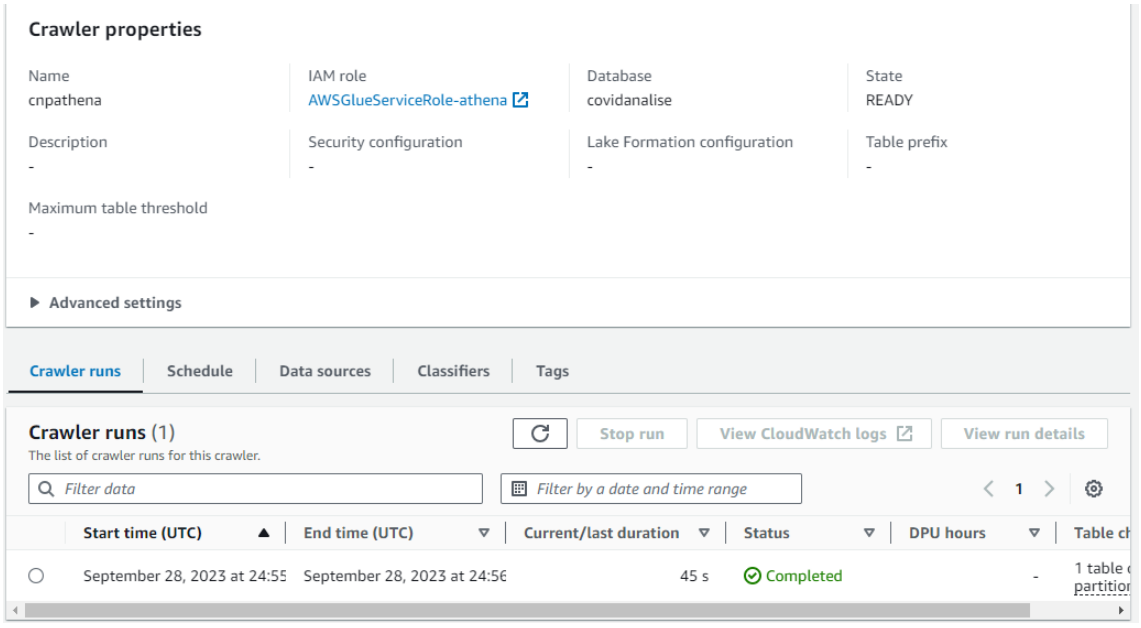
Início do processo de configuração para criar o crawler



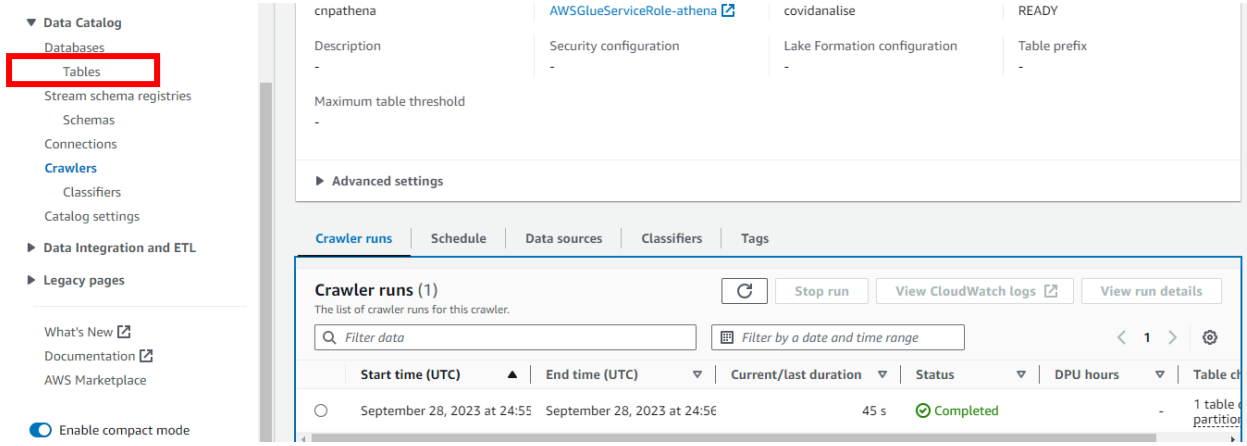
Nomear o IAM para habilitar a comunicação entre os serviços, ou seja, com o S3



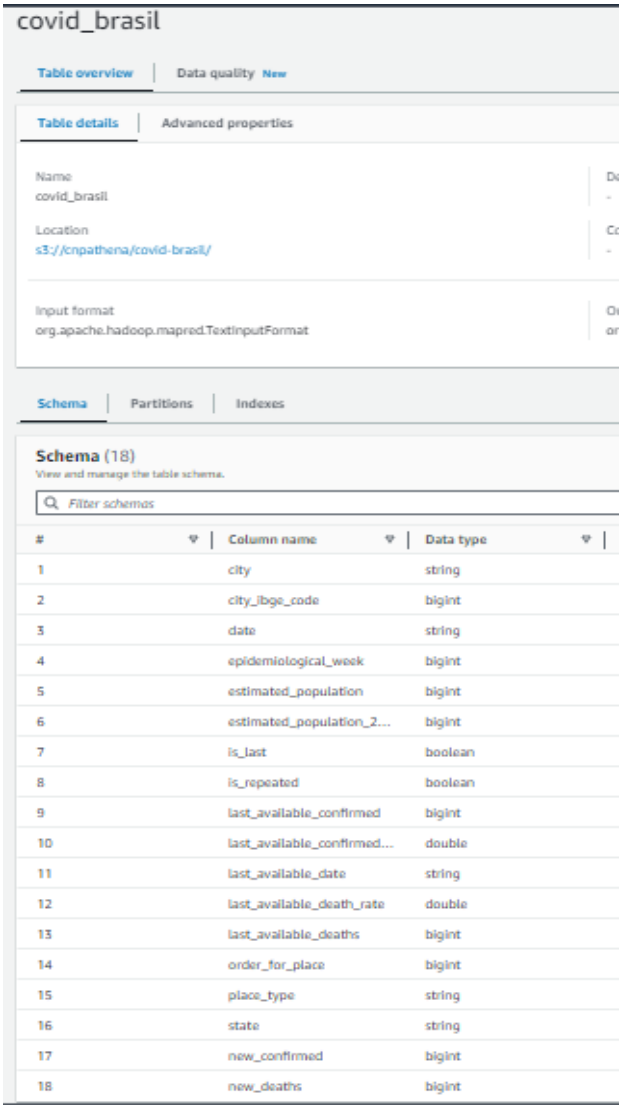
Tela mostrando que o crawler foi criado com sucesso



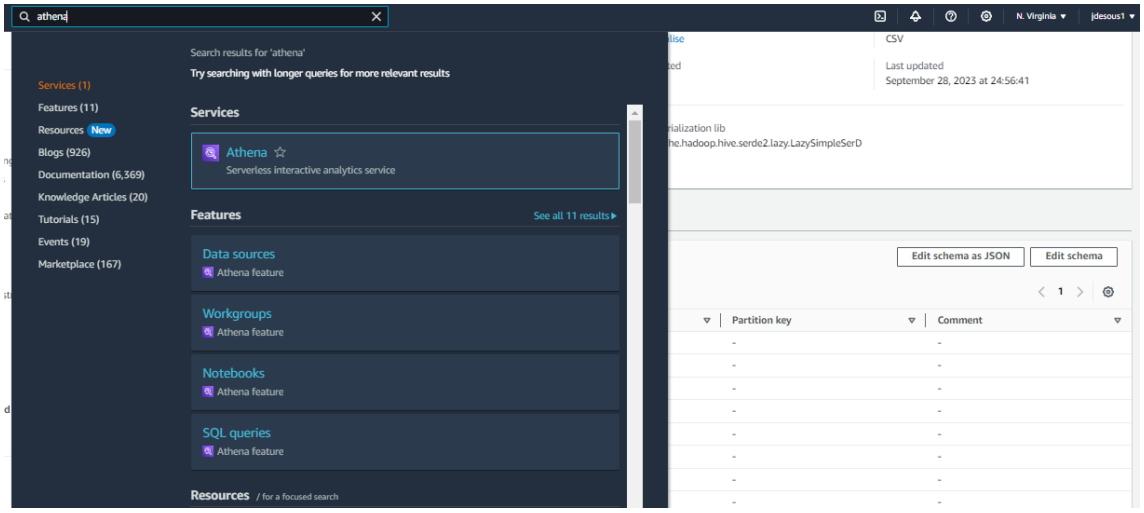
Após clicar em “Run Crawler” o arquivo é executado no mesmo



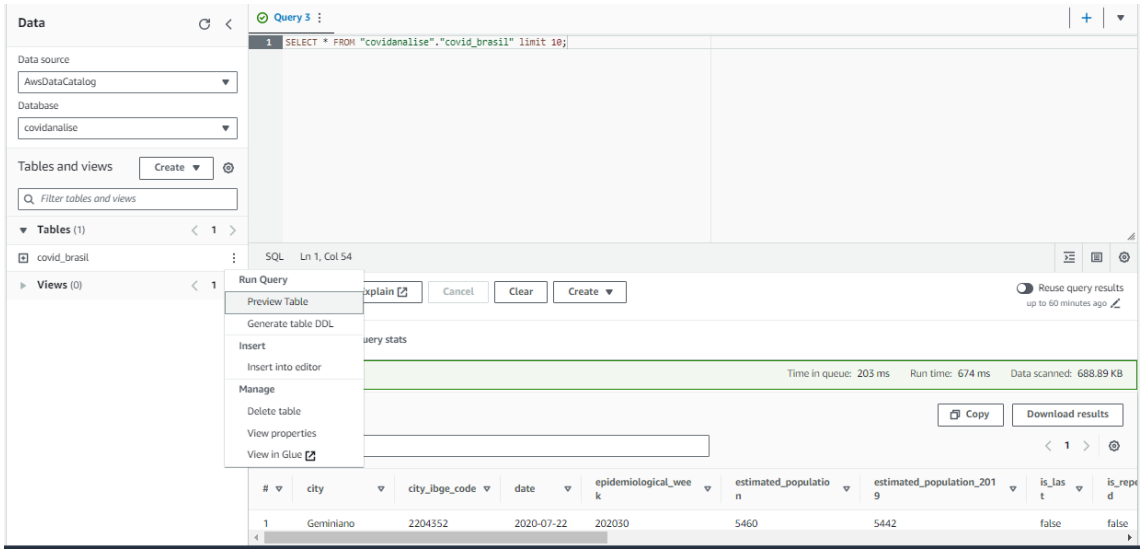
Passo seguinte é clicar em tabelas



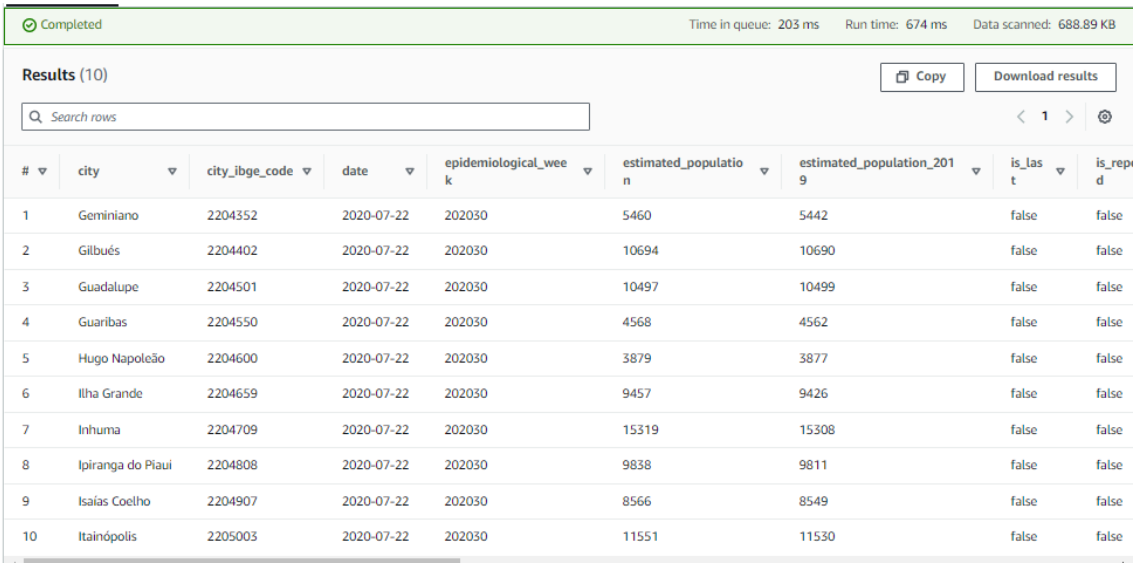
O esquema mostrado apresenta as colunas presentes no arquivo e os tipos de variáveis contidos nele



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

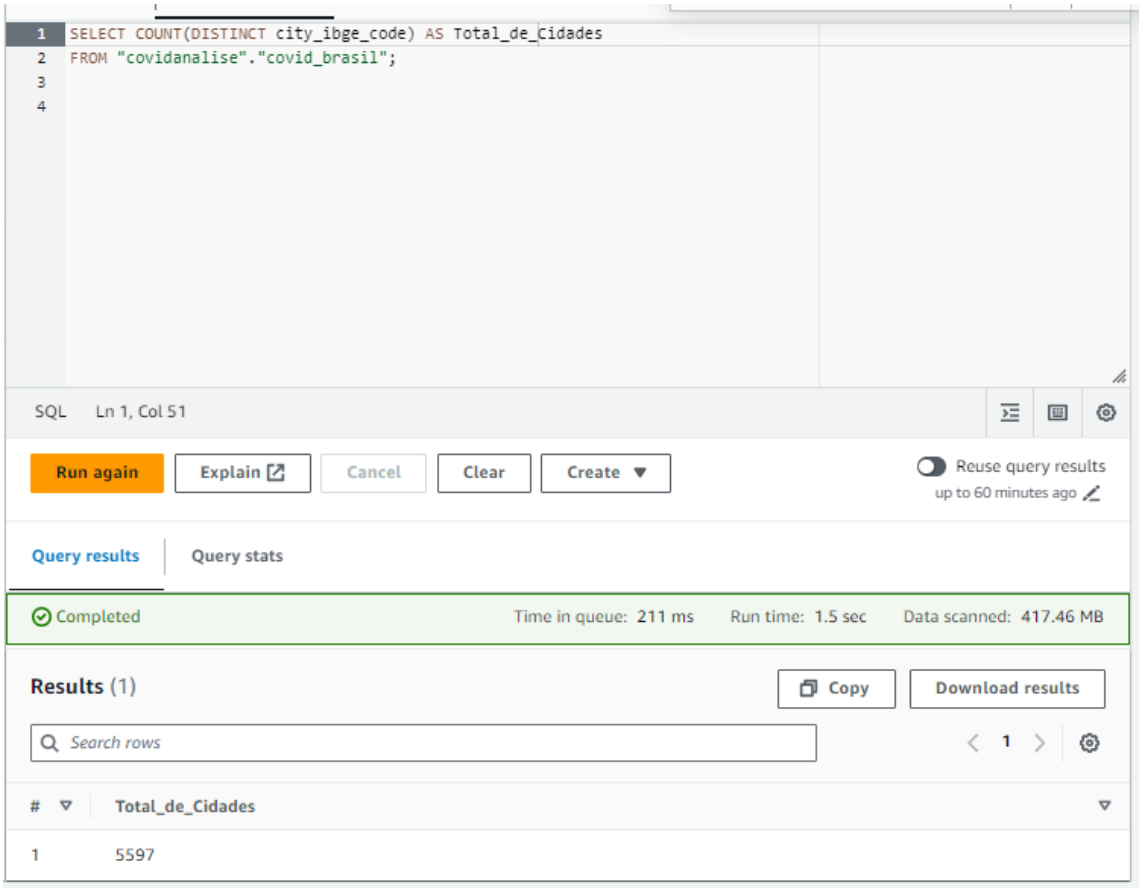
Search rows — 1

#	city	city_ibge_code	date	epidemiological_week	estimated_population	estimated_population_2019	is_last	is_repd
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:



```
1  SELECT COUNT(DISTINCT city_ibge_code) AS Total_de_Cidades
2  FROM "covidanalise"."covid_brasil";
3
4
```

SQL Ln 1, Col 51

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

Query results | Query stats

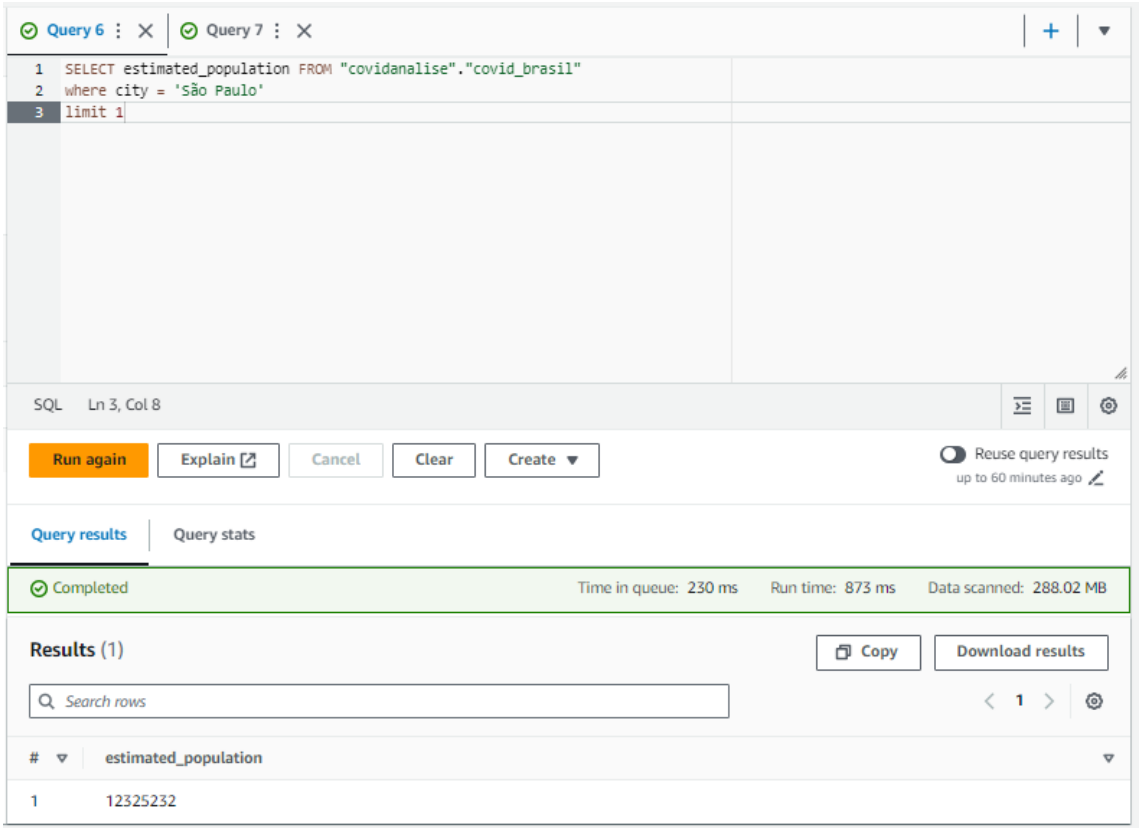
Completed — Time in queue: 211 ms — Run time: 1.5 sec — Data scanned: 417.46 MB

Results (1) — Copy — Download results

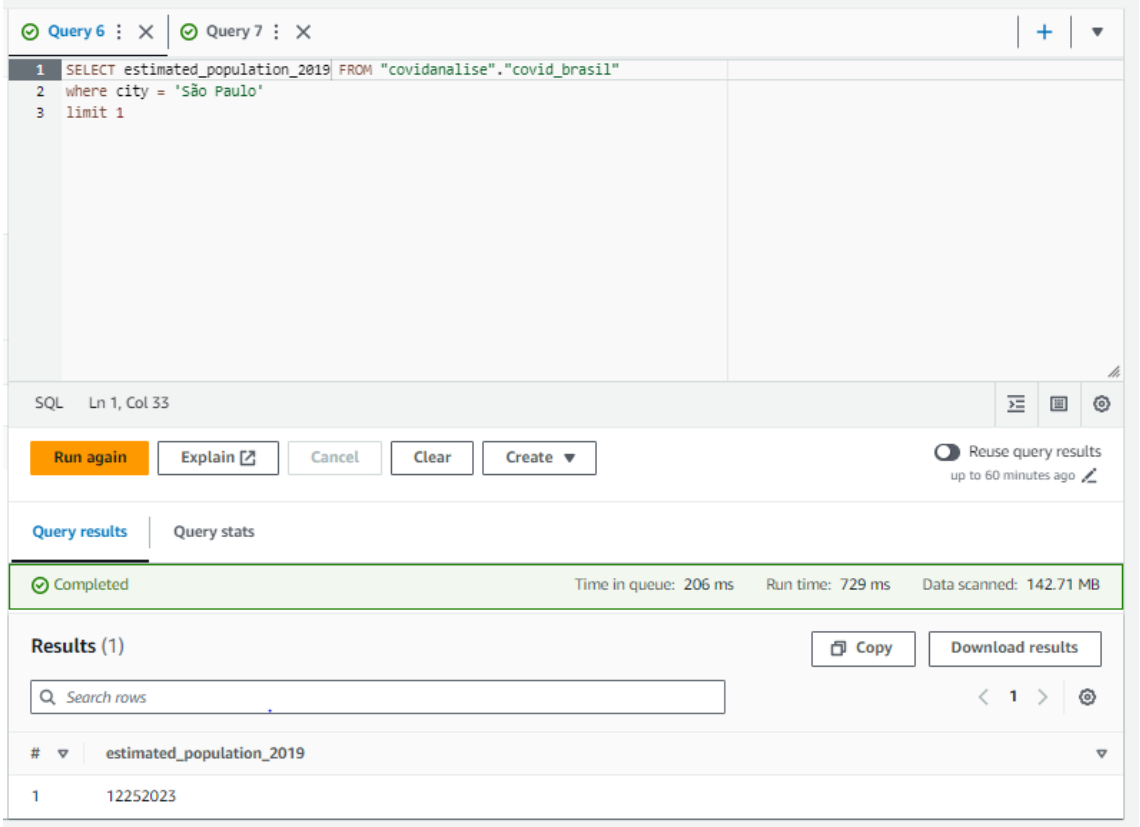
Search rows — 1

#	Total_de_Cidades
1	5597

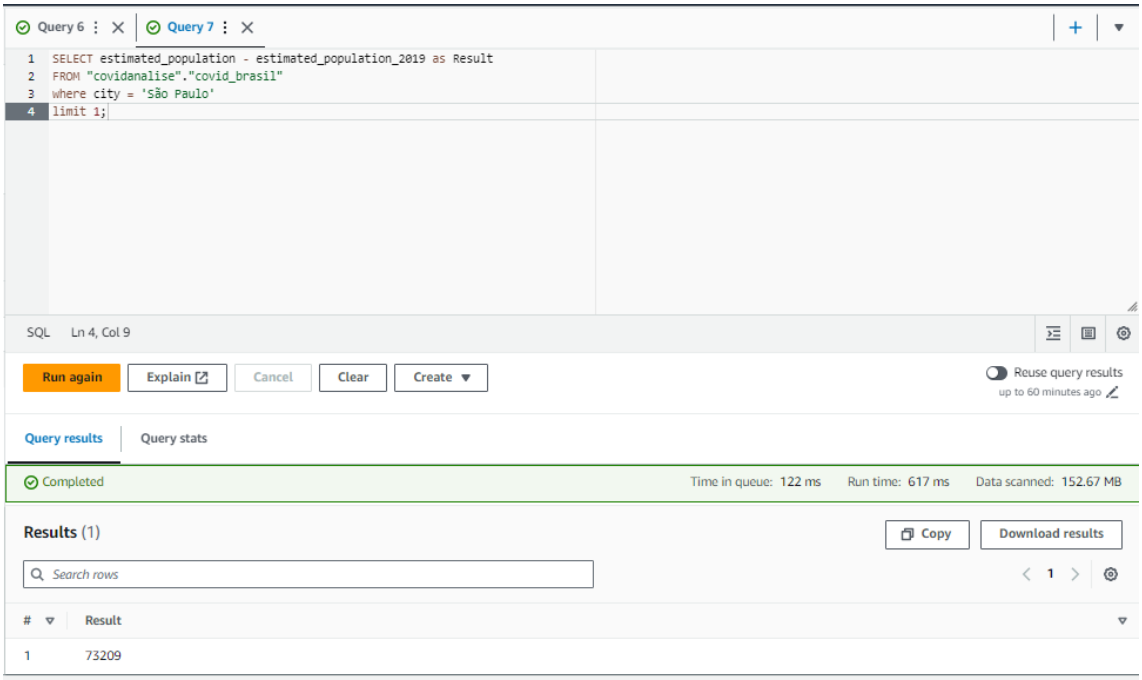
Total de cidades afetadas pela Covid 19 com base na tabela exploradas



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



A população estimada em 2019 era de 12.252.023 habitantes



Total de mortes confirmadas: 73209