

Trabalhando os dados no ASTRA (Cassandra)

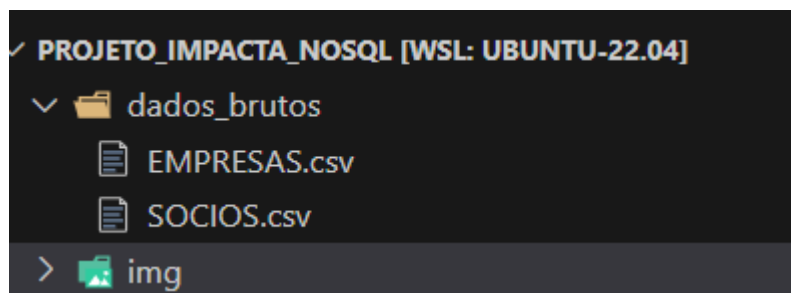
Nesta sessão vamos manipular os dados dentro do serviço do **DATASTAX**.

Sobre os dados que vamos utilizar:

Utilizaremos os arquivos **EMPRESAS.csv** e **SOCIOS.csv** do site **GOV.BR**

- EMPRESAS: Uma amostra de registros contendo o nome da empresa e seu capital social.
- SOCIOS: Uma amostra de registros contendo o nome dos socios de uma determinada empresa e sua data de entrada.

Foi criado na raíz do projeto uma pasta chamada "dados_brutos", onde deixamos guardados estes arquivos.



O que vamos fazer com os dados:

Primeiramente vamos analisar os dados dos dois arquivos separadamente e responder algumas perguntas sobre eles:

1. EMPRESAS:

- Qual o nome da empresa que possui o maior capital social?

2. SOCIOS:

- Qual o nome do Socio que está a mais tempo numa determinada empresa?

Depois vamos juntar os dois arquivos e responder outras questões:

- Quais são os nomes dos Sócios da empresa que possui o maior capital social?
- Qual o nome da empresa que possui o socio mais antigo?

Sobre a carga dos dados:

Neste passo optamos por inclui os dados na plataforma via driver do ASTRA, segue abaixo trecho de código do qual inicializamos uma sessão:

```
from astrapy.db import AstraDB
from dotenv import load_dotenv

import os

load_dotenv()
```

```
# Initialization
db = AstraDB(
    token=os.environ.get("ASTRA_TOKEN2"),
    api_endpoint="https://8831b4f4-a967-4440-922a-5ea289a1c5e6-us-east-1.apps.astra.datastax.com")

print(f"Connected to Astra DB: {db.get_collections()}")
```

Precisamos agora transferir nossos dados, porém antes de tudo vamos criar uma coleção chamada **empresas**, depois vamos transformar nossos arquivos .csv em listas de dicionários para por fim inclui-los.

As inclusões vão ser feitas dentro da coleção criada. Uma vez feita as inclusões, os dados estarão dentro do sistema **DATASTAX**.

```
import pandas as pd

collection = db.create_collection("empresas")

df_empresas = pd.read_csv('../dados_brutos/EMPRESAS.csv', sep=';')
df_socios = pd.read_csv('../dados_brutos/SOCIOS.csv', sep=';')

df_inner_join =
pd.merge(df_empresas[['CNPJ_BASICO', 'RAZAO_SOCIAL', 'CAPITAL_SOCIAL']], \

df_socios[['CNPJ_BASICO', 'NOME', 'DATA_ENTRADA_SOCIEDADE']], \
        how="inner", on="CNPJ_BASICO").dropna()

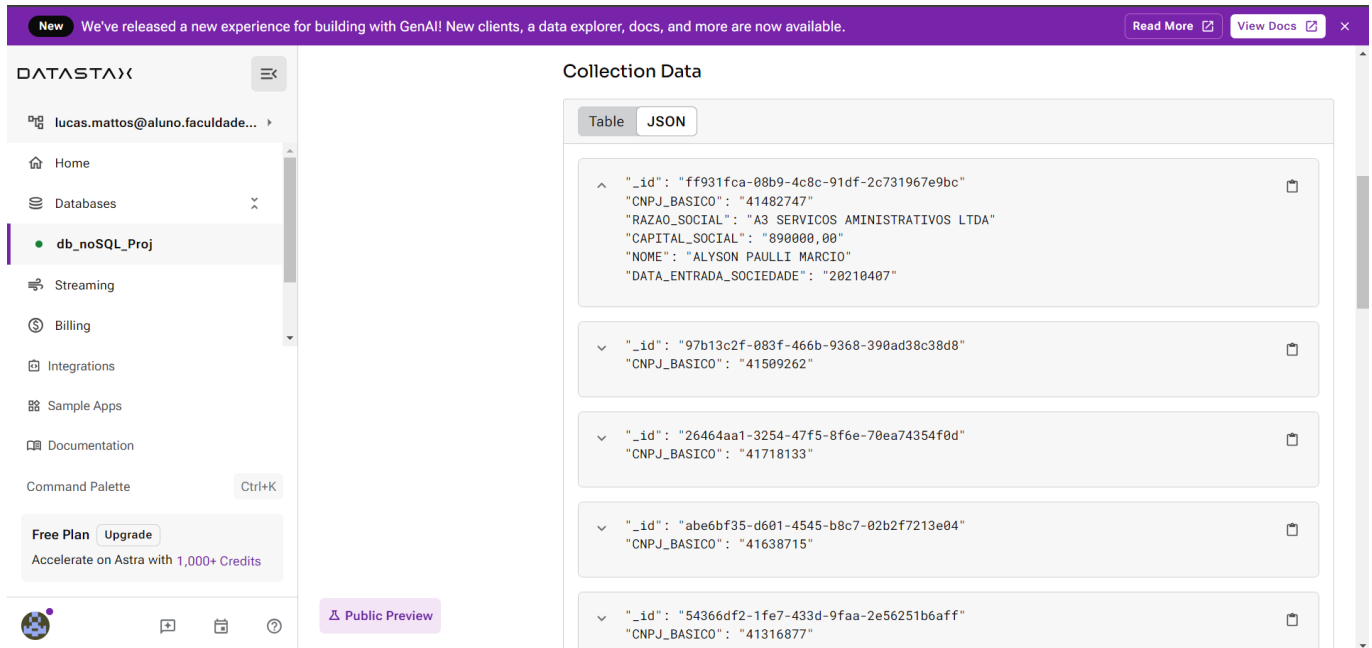
dict_of_dicts_empresas = df_inner_join.T.to_dict()

empresas = list(dict_of_dicts_empresas.values())

for empresa in empresas:
    collection.insert_one(empresa)
```

Obs: reparem que por praticidade já incluímos os dois arquivos juntos para dentro da plataforma utilizando o método **merge** da biblioteca pandas.

Feita as inclusões, os dados irão aparecer na página do DATASTAX desta forma abaixo:



The screenshot shows the DATASTAX web interface. On the left is a sidebar with navigation links: Home, Databases, db_noSQL_Proj (selected), Streaming, Billing, Integrations, Sample Apps, Documentation, and Command Palette (Ctrl+K). At the bottom of the sidebar is a 'Free Plan' badge with an 'Upgrade' button and text 'Accelerate on Astra with 1,000+ Credits'. The main area is titled 'Collection Data' and shows a list of JSON documents. The first document is expanded, showing fields: '_id', 'CNPJ_BASICO', 'RAZAO_SOCIAL', 'CAPITAL_SOCIAL', 'NOME', and 'DATA_ENTRADA_SOCIEDADE'. Below it are four more documents, each with '_id' and 'CNPJ_BASICO' fields. A 'Public Preview' button is at the bottom left of the main area.

Document	_id	CNPJ_BASICO	RAZAO_SOCIAL	CAPITAL_SOCIAL	NOME	DATA_ENTRADA_SOCIEDADE
1	"ff931fca-08b9-4c8c-91df-2c731967e9bc"	"41482747"	"A3 SERVICOS ADMINISTRATIVOS LTDA"	"890000,00"	"ALYSON PAULLI MARCIO"	"20210407"
2	"97b13c2f-083f-466b-9368-390ad38c38d8"	"41509262"				
3	"26464aa1-3254-47f5-8f6e-70ea74354f0d"	"41718133"				
4	"abe6bf35-d601-4545-b8c7-02b2f7213e04"	"41638715"				
5	"54366df2-1fe7-433d-9faa-2e56251b6aff"	"41316877"				

Visualização os dados dentro do astra:

Partimos para a fase de responder as questões propostas acima, utilizaremos os comandos **Cassandra Query Language (CQL)** disponibilizado dentro do DATASTAX.

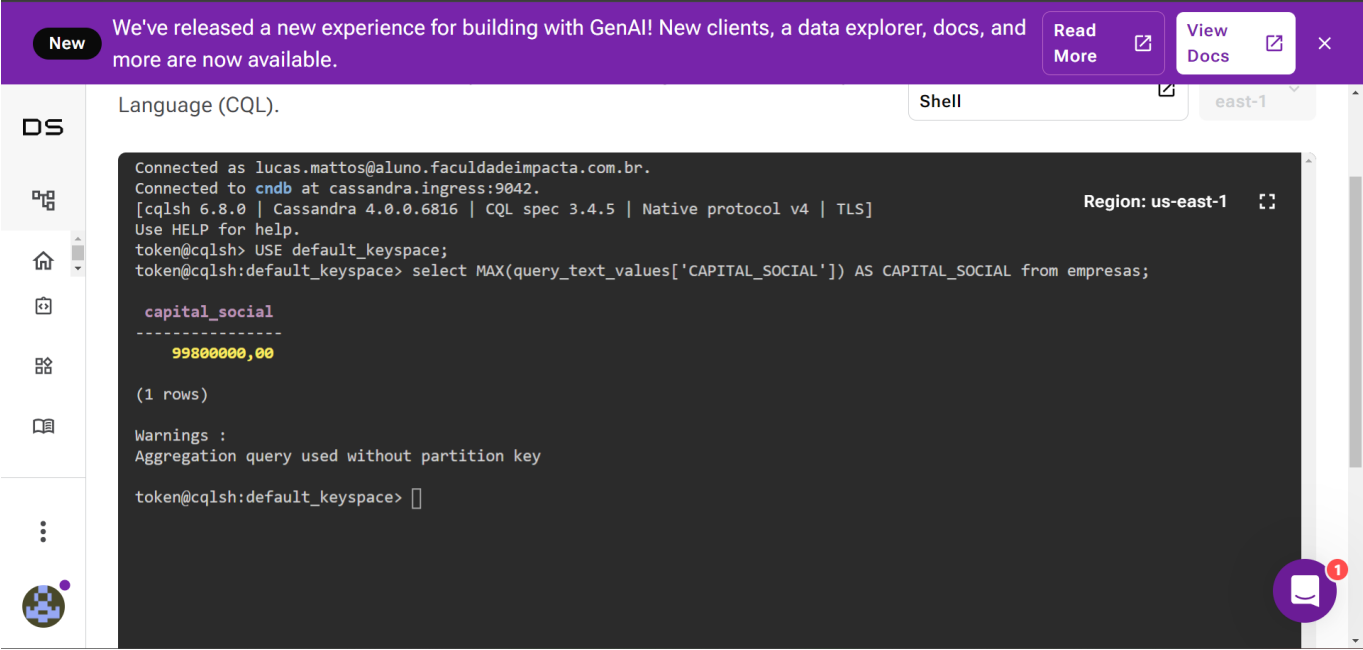
É importante ressaltar que nossa collection foi criada dentro de uma database padrão da plataforma (default_keyspace). Sempre que formos explorar nossos dados temos sempre que direcionar a database da qual pertence por meio do comando **USE default_keyspace;**

1. Qual o nome da empresa que possui o maior capital social?

Para chegarmos ao resultado primeiro devemos criar uma query que retorne o valor máximo do capital social da amostra, em seguida usar o resultado encontrado como filtro para encontrar a empresa correspondente. Segue abaixo os comandos e print dos resultados:

(1)

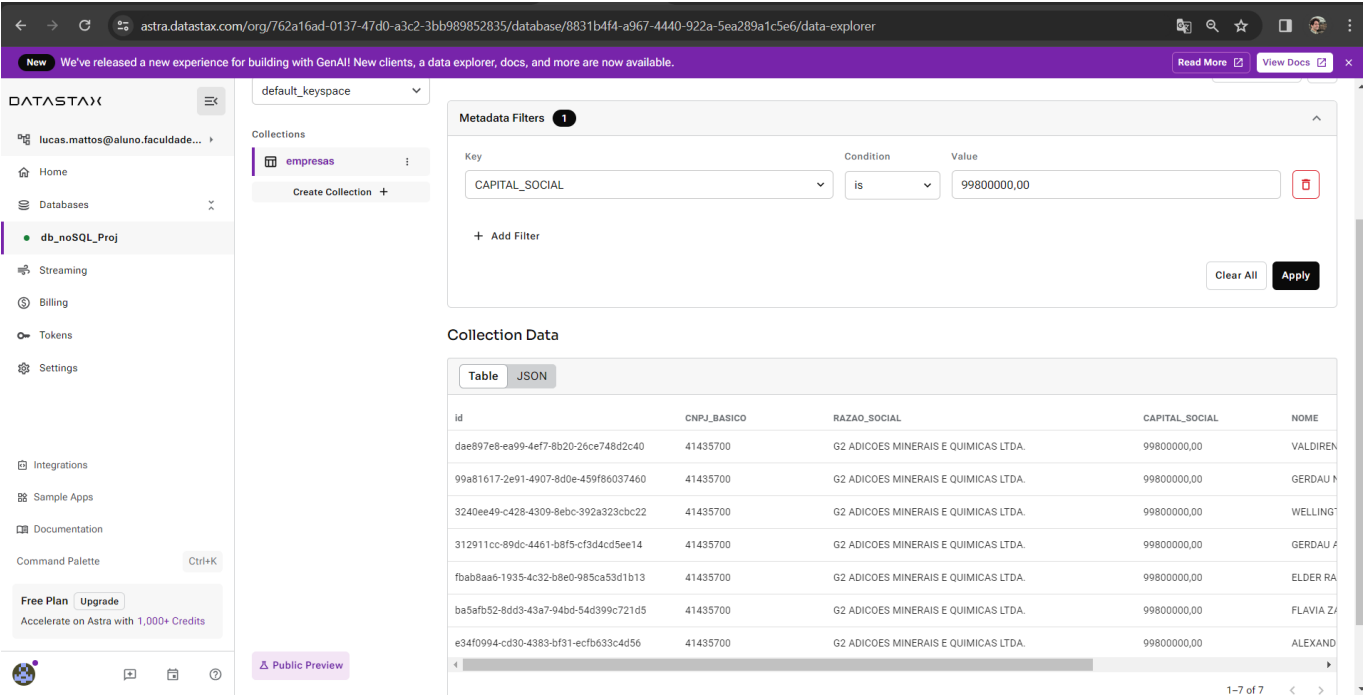
```
select MAX(query_text_values['CAPITAL_SOCIAL']) AS CAPITAL_SOCIAL from empresas;
```



O maior capital social corresponde ao valor de R\$ 99.800.000,00.

(2)

Incluindo o valor da query acima no data explorer do DATASTAX.



Portanto podemos afirmar que a empresa **G2 ADICOES MINERAIS E QUIMICAS LTDA.** é a que possui o maior capital social.

2. Qual o nome do Socio que está a mais tempo numa determinada empresa?

Podemos encontrar essa informação no campo DATA_ENTRADA_SOCIEDADE. Quanto mais antiga for a data, maior é o tempo que o socio está vinculado com a empresa. Podemos saber a data mais antiga por meio da função `MIN()`. Segue comando utilizado e print do resultado:

(1)

```
select MIN(query_dbl_values['DATA_ENTRADA_SOCIEDADE']) AS MIN_DATE from empresas;
```

The screenshot shows the Databricks interface. On the left is a sidebar with navigation icons. The main area displays a terminal window with the following content:
Connected as lucas.mattos@aluno.faculdadeimpacta.com.br.
Connected to cndb at cassandra.ingress:9042.
[cqlsh 6.8.0 | Cassandra 4.0.0.6816 | CQL spec 3.4.5 | Native protocol v4 | TLS] Region: us-east-1
Use HELP for help.
token@cqlsh> USE default_keyspace;
token@cqlsh:default_keyspace> select MIN(query_db1_values['DATA_ENTRADA_SOCIEDADE']) AS MIN_DATE from empresas;

min_date

20070820

(1 rows)

Warnings :
Aggregation query used without partition key

token@cqlsh:default_keyspace>

O socio mais antigo entrou na empresa no dia 20 de agosto de 2007.

(2)

Incluindo o valor da query acima no data explorer do DATASTAX.

New

We've released a new experience for GenAI! New clients, a data explorer, docs, and more are now available.

Read More

View Docs

DS

empresas

Create Collection +

+ Add Filter

Metadata Filters 1

Key

DATA_ENTRADA_SOCIEDADE

Condition

is

Value

20070820

+ Add Filter

Clear All

Apply

Collection Data

Table

JSON

RAZAO_SOCIAL	CAPITAL_SOCIAL	NOME	DATA_ENTRADA_SOCIEDADE
IGREJA RESTAURANDO VIDAS, MINISTERIO EBENEZER	0,00	GILDOBERTO GONCALVES BORGES	20070820

1-1 of 1

Concluimos que **GILDOBERTO GONCALVES BORGES** é o socio mais antigo da amostra.

3. Quais são os nomes dos Sócios da empresa que possui o maior capital social?

Como já juntamos as duas tabelas logo no processo de carga dos dados para o DATASTAX, podemos solucionar essa questão utilizando a resolução da primeira questão.

Logo se utilizarmos o comando `select MAX(query_text_values['CAPITAL_SOCIAL']) AS CAPITAL_SOCIAL from empresas;` e incluir o resultado (99800000,00) no filter no data explorer teremos o seguinte resultado:

DS

Collections

empresas

Create Collection +

Key

CAPITAL_SOCIAL

Condition

is

Value

99800000,00

+ Add Filter

Clear All

Apply

Collection Data

Table

JSON

	CNPJ_BASICO	RAZAO_SOCIAL	CAPITAL_SOCIAL	NOME	DATA_ENTRADA_SOCIEDADE
c40	41435700	G2 ADICOES MINERAIS E QUIMICAS LTDA.	99800000,00	VALDIRENE SULLAS TEIXEIRA PERESSINOT...	20220825
460	41435700	G2 ADICOES MINERAIS E QUIMICAS LTDA.	99800000,00	GERDAU NEXT S.A.	20220118
c22	41435700	G2 ADICOES MINERAIS E QUIMICAS LTDA.	99800000,00	WELLINGTON ARAUJO DE OLIVEIRA	20230530
y14	41435700	G2 ADICOES MINERAIS E QUIMICAS LTDA.	99800000,00	GERDAU ACOS LONGOS S.A.	20210401
b13	41435700	G2 ADICOES MINERAIS E QUIMICAS LTDA.	99800000,00	ELDER RAPACHI	20210923
1d5	41435700	G2 ADICOES MINERAIS E QUIMICAS LTDA.	99800000,00	FLAVIA ZANGRANDI MARCONDES	20221103
56	41435700	G2 ADICOES MINERAIS E QUIMICAS LTDA.	99800000,00	ALEXANDRE DE TOLEDO CORREA	20210401

1-7 of 7

Os sócios da empresa de maior capital social da amostra são: *VALDIRENE SULLAS TEIXEIRA, GERDAU NEXT S.A., WELLINGTON ARAUJO DE OLIVEIRA, GERDAU ACOS LONGOS S.A., ELDER RAPACHI, FLAVIA ZANGRANDI MARCONDES*e *ALEXANDRE DE TOLEDO CORREA*.

4. Qual o nome da empresa que possui o socio mais antigo?

O mesmo dito na questão anterior se aplica nesse caso. Como os dados de empresa e socios então todos juntos na mesma tabela, uma vez encontrado o socio mais antigo também encontramos a empresa da qual ele faz parte.

DS

Collections

empresas

Create Collection +

Metadata Filters 1

Key

DATA_ENTRADA_SOCIEDADE

Condition

is

Value

20070820

+ Add Filter

Clear All

Apply

Collection Data

Table

JSON

	RAZAO_SOCIAL	CAPITAL_SOCIAL	NOME	DATA_ENTRADA_SOCIEDADE
	IGREJA RESTAURANDO VIDAS, MINISTERIO EBENEZER	0,00	GILDOBERTO GONCALVES BORGES	20070820

1-1 of 1

Portanto a empresa que possui o socio mais antigo é a *IGREJA RESTAURANDO VIDAS, MINISTERIO EBENEZER*.