

MVP – Disciplina: Sprint Engenharia de Dados

Aluna: Amanda Lins Guerra

TEMA: EMPRESAS ATIVAS DA CIDADE DO RECIFE - PE

Fonte de dados: Secretaria de Finanças da Prefeitura do Recife

Link de conjuntos de dados: [Empresas da Cidade do Recife - Conjuntos de dados - Portal de Dados Abertos da Cidade do Recife](#)



## Introdução

O referente projeto se baseia no conjunto de dados que contêm a descrição das Empresas da Cidade do Recife com os seus respectivos endereços e atividades, que estão inscritas como contribuintes no Município do Recife, Estado de Pernambuco.

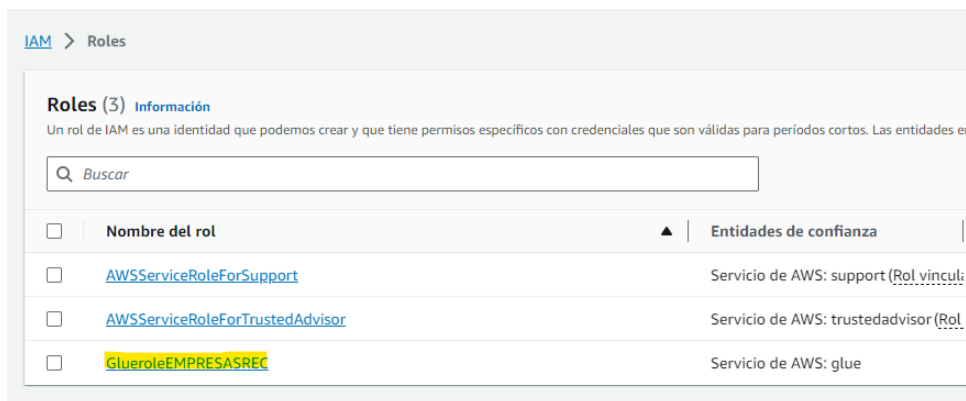
## Ferramentas utilizadas para ETL e Queries

Para fazer o upload e armazenar os buckets dos dados utilizei o **Amazon S3** que é um serviço de armazenamento de objetos. Para realizar ETL do conjunto de dados utilizei o **AWS Glue** que é um serviço de integração de dados com tecnologia sem servidor com pipelines visuais. Para execucao de queries utilizei o **Amazon Athena** é um serviço de consultas interativas que facilita a análise de dados do Amazon S3 usando **SQL** padrão.

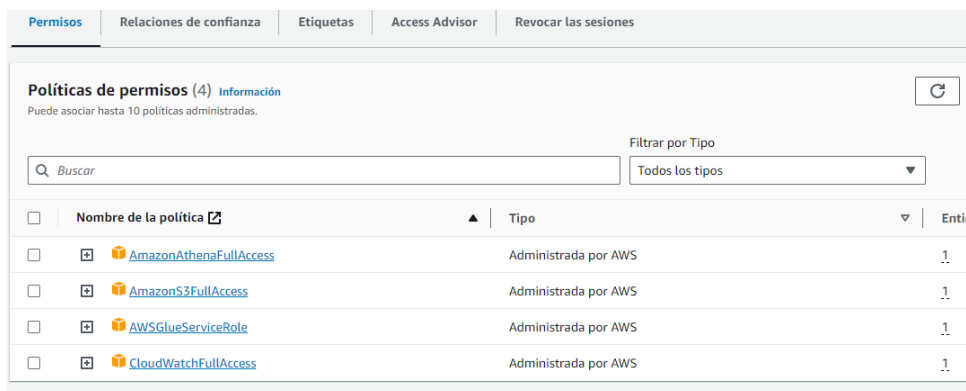
## Passo a passo do projeto

### 1 – Criação de Role e Configuração das permissões:

Criei primeiro o Role para poder utilizar o serviço Glue. O Rol se chamou “GlueroleEMPRESASREC”:



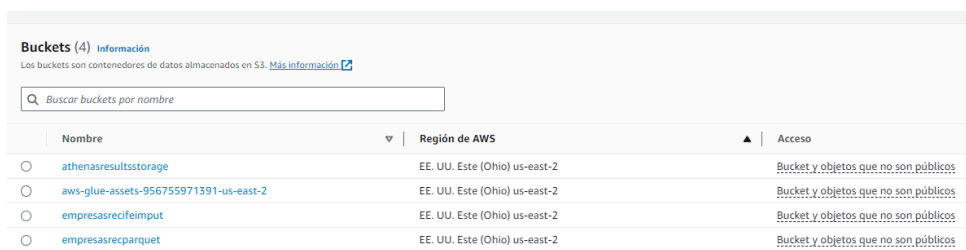
Em seguida configurei as permissões necessárias para utilizar S3, AWS Glue e Athena:



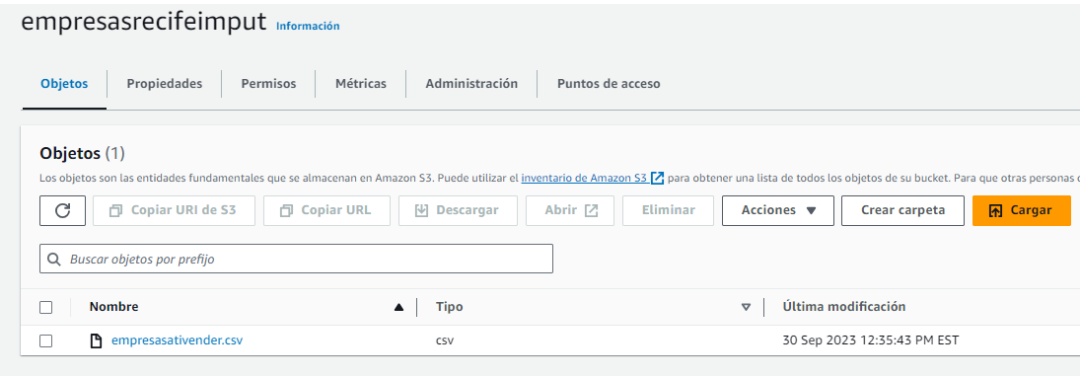
### 2 – Criação e alimentação de Buckets in S3

Criei 3 buckets:

- Empresasrecimput (input de arquivo .CSV)
- Empresasrecparquet (output arquivo formato parquet)
- Ethenasresultsstorage (bucket para armazenamento de queries de Athena)



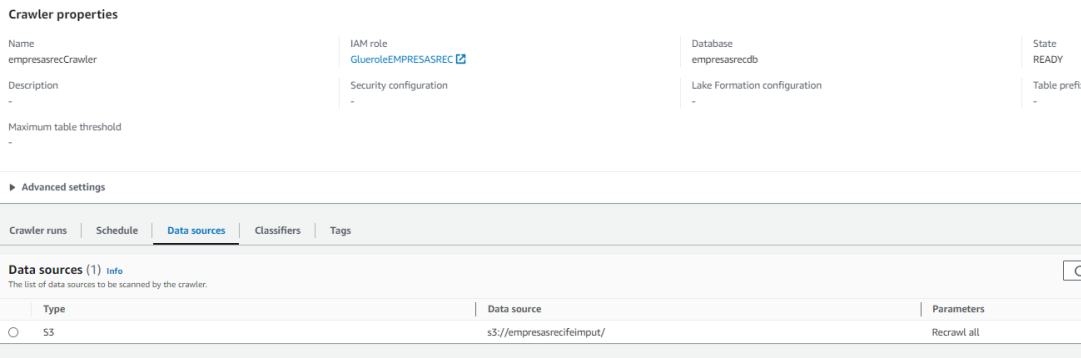
3 – Upload de arquivo empresasativender.csv ao bucket empresasrecimput:



4 – Criacao de Crawler para arquivo .csv:

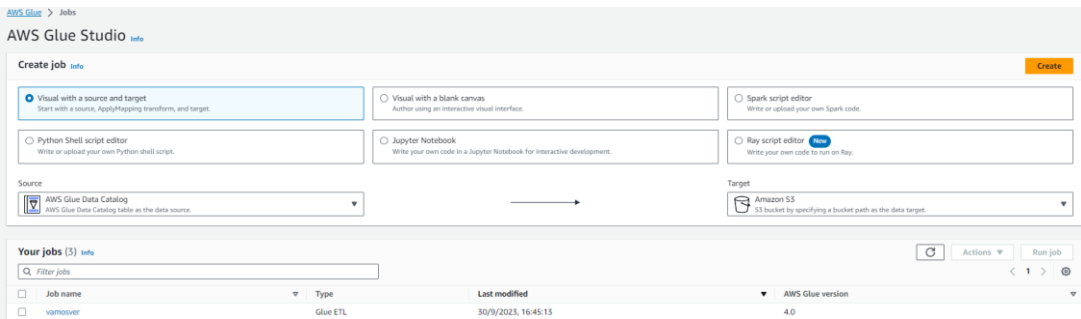
A criação do crawler é necessária para inspeção do arquivo como um todo, para classificar, agrupar dados em tabelas ou partições e gravar metadados no Data Catalog.

Criei o crawler para inspeção do arquivo imput do bucket “empresasrecimput”:



5 – Criacao de ETL JOB:

Criei o ETL JOB de nome “vamosver”, com source **AWS Data Catalog** e target Amazon S3.



6 – Criação de pipeline visual de ETL:

Primeiro defini a DB “empresasrecdb” e Table “empresasrecifeimput”:

**Data source properties - Data Catalog**

Name  
AWS Glue Data Catalog

Database  
Choose a database.  
empresasrecdb

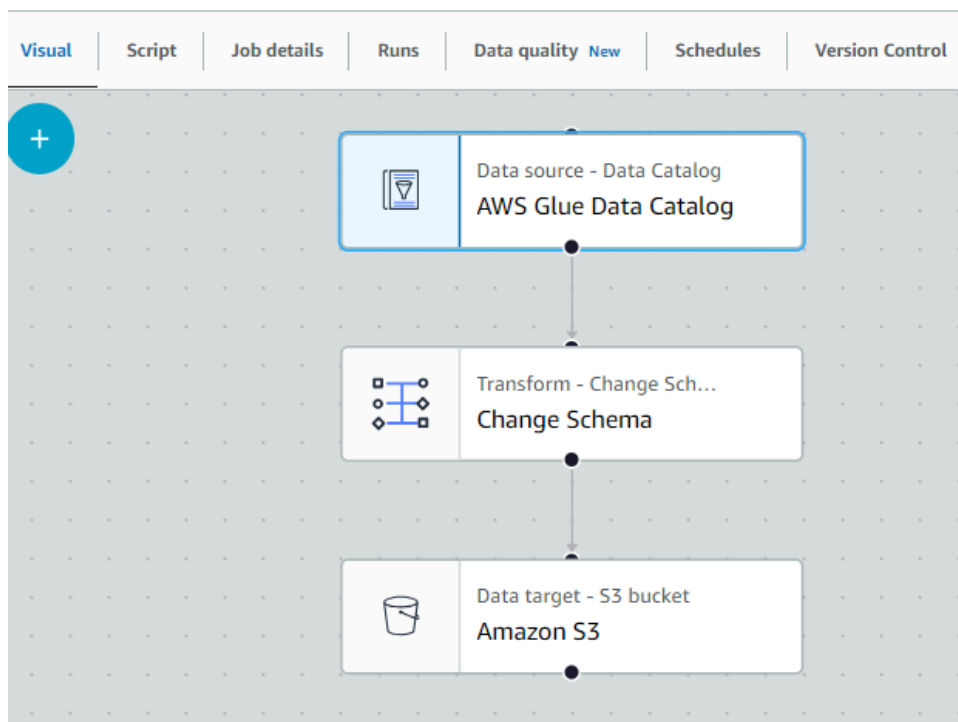
► Use runtime parameters

Table  
empresasrecifeinput

► Use runtime parameters

Em seguida no pipeline coloquei um passo de transformação “Change Schema”:

**vamosver**



Para logo fazer os “drops” das colunas que me pareciam dispensáveis:

Transform	Output schema	Data preview	
Change Schema (Apply mapping)			
Source key	Target key	Data type	Drop
razao_social	<input type="text" value="razao_social"/>	<input type="text" value="string"/>	<input type="checkbox"/>
nome_fantasia	<input type="text" value="nome_fantasia"/>	<input type="text" value="string"/>	<input type="checkbox"/>
cod_logradouro	<input type="text" value="cod_logradouro"/>	<input type="text" value="long"/>	<input type="checkbox"/>
nome_logradouro	<input type="text" value="nome_logradouro"/>	<input type="text" value="string"/>	<input type="checkbox"/>
numero_residencia	<input type="text" value="numero_residencia"/>	<input type="text" value="string"/>	<input type="checkbox"/>
numero_lote			<input checked="" type="checkbox"/>
cod_bairro	<input type="text" value="cod_bairro"/>	<input type="text" value="long"/>	<input type="checkbox"/>
nome_bairro	<input type="text" value="nome_bairro"/>	<input type="text" value="string"/>	<input type="checkbox"/>
situacao_empresa	<input type="text" value="situacao_empresa"/>	<input type="text" value="string"/>	<input type="checkbox"/>
data_abertura_empresa	<input type="text" value="data_abertura_empresa"/>	<input type="text" value="string"/>	<input type="checkbox"/>
data_encerramento			<input checked="" type="checkbox"/>
cod_grupo	<input type="text" value="cod_grupo"/>	<input type="text" value="long"/>	<input type="checkbox"/>
nome_grupo	<input type="text" value="nome_grupo"/>	<input type="text" value="string"/>	<input type="checkbox"/>
cnae	<input type="text" value="cnae"/>	<input type="text" value="long"/>	<input type="checkbox"/>
desc_atividade	<input type="text" value="desc_atividade"/>	<input type="text" value="string"/>	<input type="checkbox"/>
atividade_principal	<input type="text" value="atividade_principal"/>	<input type="text" value="string"/>	<input type="checkbox"/>
atividade_vig_sanitaria	<input type="text" value="atividade_vig_sanitaria"/>	<input type="text" value="string"/>	<input type="checkbox"/>
atividade_predominante	<input type="text" value="atividade_predominante"/>	<input type="text" value="string"/>	<input type="checkbox"/>
incomodo	<input type="text" value="incomodo"/>	<input type="text" value="string"/>	<input type="checkbox"/>
latitude			<input checked="" type="checkbox"/>
longitude			<input checked="" type="checkbox"/>

## 7 – Definicao de output:

Defini o formato do output em parquet, com compressao snappy para ocupar menos espaco de armazenamento.

Coloquei como output uma tabela chamada “results”.

Data target properties - S3	Output schema	Data preview
-----------------------------	---------------	--------------

Name

Amazon S3

Node parents

Choose which nodes will provide inputs for this one.

Choose one or more parent node

Change Schema

ApplyMapping - Transform

Format

Parquet

Compression Type

Snappy

S3 Target Location

Choose an S3 location in the format s3://bucket/prefix/object/ with a trailing slash (/).

s3://empresasrecparquet/

Data Catalog update options

Choose how you want to update the Data Catalog table's schema and partitions. These options will only apply if the Data Catalog table is an S3 backed source.

☐ Do not update the Data Catalog

☒ Create a table in the Data Catalog and on subsequent runs, update the schema and add new partitions

☐ Create a table in the Data Catalog and on subsequent runs, keep existing schema and add new partitions

Database

Choose the database from the AWS Glue Data Catalog.

empresasrecdb

► Use runtime parameters

Table name

Enter a table name for the AWS Glue Data Catalog.

Results

## 8 – Execução de ETL JOB

Job runs (13) [info](#)

Filter job runs by property

Job name	Type	Start time	End time	Run status	Run time	Capacity	Worker type	DPU hours
vamosver	Glue ETL	09/30/2023 16:45:30	09/30/2023 16:46:58	Succeeded	1 minute	10	G.1X	0.22

## 9 – Execução de crawler para DB

Após transformação da tabela csv, tive que rodar o crawler sobre a DB para inspecao da tabela csv modificada.

vamosver

**Crawler properties**

Name	vamosver	IAM role	<a href="#">GlueRoleEMPRESASREC</a>	State	READY
Security configuration	-	Lake Formation configuration	-		

► Advanced settings

Crawler runs | Schedule | **Glue tables** | Classifiers | Tags

**Glue tables (1)** [info](#)

The list of AWS Glue Catalog tables to be scanned by the crawler.

Database	Tables
empresasrecdb	1 table

## 10 – Iniciar a aplicacao Athena:

Executei o Athena e primeiro cliquei em atualizar Data, então a minha tabela target output “results” apareceu em Tables, já com todos os nomes de colunas atualizados.

**Data**

Data source: [AwsDataCatalog](#)

Database: [empresasrecdb](#)

Tables and views: [Create](#)

Search: results

**Tables (2)**

- results
- cnnpj
- razao\_social
- nome\_fantasia
- cod\_logradouro
- nome\_logradouro
- numero\_residencia
- cod\_bairro
- nome\_bairro
- situacao\_empresa
- data\_abertura\_empresa
- cod\_grupo
- nome\_grupo
- cnae
- desc\_atividade
- atividade\_principal
- atividade\_vig\_sanitaria

**Views (0)**

# Execução de *QUERIES* (consultas) em Amazon Athena

## 1 - Qualidade dos dados

Minhas primeiras consultas foram para descartar dados faltantes. Através dos queries, comprovei que não havia problemas no conjunto de dados nesse sentido; não encontrei campos nulos em CNPJ, que é a chave principal, e tampouco havia campos nulos em Razão \_Social.

Queries:

```
SELECT cnpj  
  
FROM results_new  
  
WHERE cnpj IS NULL
```

```
SELECT razao_social  
  
FROM results_new  
  
WHERE razao_social IS NULL
```

## 2 – Informações básicas das empresas

Pelas consultas constatei que atualmente há 317572 empresas ativas na cidade do Recife. As empresas são classificadas em 29 áreas de atuação (nome\_grupo) e dentro dessas áreas há 1232 atividades econômicas (desc\_atividades).

Toda empresa tem CNPJ, Razão Social e Nome Fantasia. E 1,17% das empresas possuem Razão Social = Nome Fantasia.

Queries:

```
SELECT count(cnpj)  
  
from results  
  
-----  
  
CREATE VIEW view_name AS  
  
SELECT razao_social, nome_fantasia  
  
FROM results  
  
WHERE razao_social = nome_fantasia;
```

### 3 – Bairros com maior incidência de empresas ativas:

Boa viagem é um dos bairros mais populosos e extensos de Recife, é também considerado bairro nobre pela praia de Boa Viagem e por ser grande polo empresarial e turístico.

É de se esperar que os dados revelem que Boa Viagem tem a maior densidade de empresas ativas (53081 empresas), seguido do Pina (21314 empresas), que é adjacente ao mesmo. Em terceiro lugar está Boa Vista (20285 empresas) que é parte do centro do Recife.



Boa Viagem e Pina:



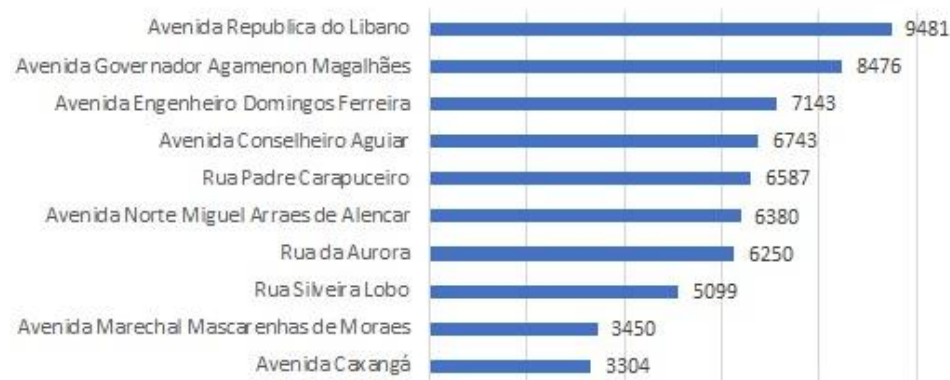
#### QUERY UTILIZADA:

```
SELECT Nome_Bairro, COUNT(CNPJ) AS Total_CNPJ
FROM Results
GROUP BY Nome_Bairro
ORDER BY Total_CNPJ DESC;
```



#### 4 – Ruas com maior incidência de empresas ativas:

Um dado interessante é ver que uma única avenida abriga 9481 empresas. É uma importante avenida do bairro do Pina.

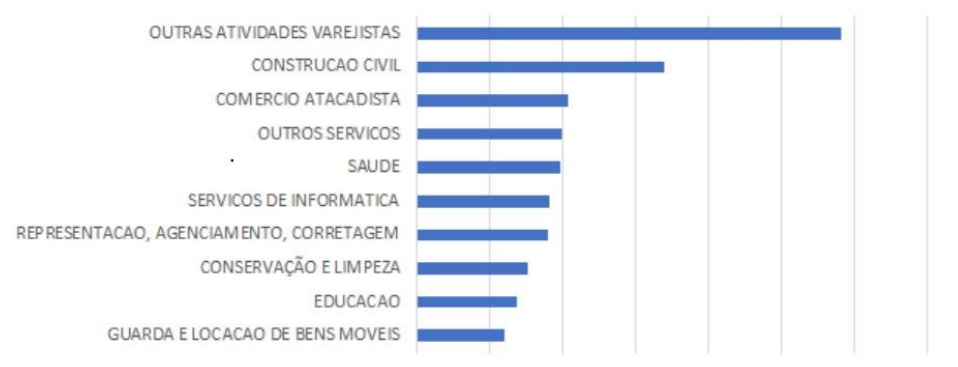


QUERY UTILIZADA:

```
SELECT Nome_logradouro, COUNT(CNPJ) AS Total_CNPJ  
FROM Results  
GROUP BY Nome_logradouro  
ORDER BY Total_CNPJ DESC;
```

#### 5 - Áreas de atuação e atividades econômicas com maior incidência:

Dentro das áreas de atuação, predominam “outras atividades varejistas” e construção civil.



E a atividade de maior incidência é de “serviços combinados de escritório e apoio administrativo”.



#### QUERIES UTILIZADAS:

```
SELECT nome_grupo, COUNT(CNPJ) AS Total_CNPJ  
FROM Results  
GROUP BY Nome_grupo  
ORDER BY Total_CNPJ DESC;
```

```
-----  
SELECT desc_atividade, COUNT(CNPJ) AS Total_CNPJ  
FROM Results  
GROUP BY desc_atividade  
ORDER BY Total_CNPJ DESC;
```

## 6 – Empresas com o maior número de áreas de atuação:

Para esse query eu consultei quais empresas tinham registradas mais de 15 áreas de atuação.



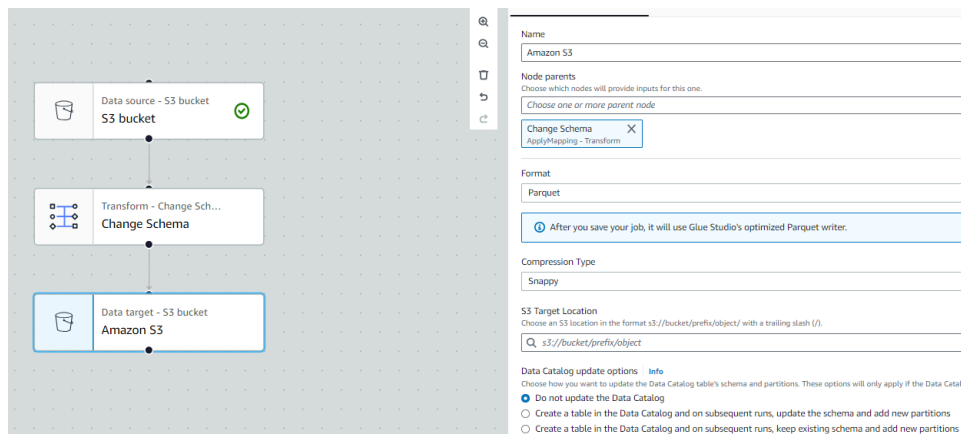
#### QUERY UTILIZADA:

```
SELECT razao_social, COUNT(DISTINCT cod_grupo) AS num_categorias  
FROM Results  
GROUP BY razao_social  
HAVING COUNT(DISTINCT cod_grupo)>15  
order by num_categorias desc
```

# Dificuldades encontradas

## 1 - Criação do Pipeline – Source Setting e Arquivo Parquet valores vazios

Durante a fase de criação de pipeline do ETL JOB, encontrei muita dificuldade porque inicialmente estava colocando o source como S3 Bucket e não como AWS Catalog.



Ao executar o ETL JOB, gerava um arquivo parquet com tamanho muito pequeno, próximo de 1kb, que resultou sendo uma tabela com valores nulls, sem qualquer conteúdo, somente com os nomes das colunas.

Depois de tentar muitas vezes, decidi tentar colocando o source como AWS Catalog e mudei a opção de Data Catalog options:

Data Catalog update options | Info  
Choose how you want to update the Data Catalog table's schema and partitions. These options will only apply if the Data Catalog table is an S3 backed source.

- ☐ Do not update the Data Catalog
- ☒ Create a table in the Data Catalog and on subsequent runs, update the schema and add new partitions
- ☐ Create a table in the Data Catalog and on subsequent runs, keep existing schema and add new partitions

Após fazer isso, funcionou e os arquivos gerados em parquet tinha tamanho mais razoável, de aprox. 5mb e todas as células da tabela tinham conteúdo.

## 2 – Função datediff

Tive a intenção de calcular o tempo de existência de cada empresa da tabela, baseando-me na data de criação das empresas, porém, não foi possível mesmo após muitas tentativas de códigos diferentes.

Algumas das tentativas de Queries:

```
SELECT
```

```
data_abertura_empresa,
```

```
CASE
```

```

        WHEN TRY_CAST(DATE_PARSE(data_abertura_empresa, 'dd/M/yyyy') AS date) IS NOT NULL THEN
'Formato válido de fecha'

        ELSE 'Formato inválido de fecha'

    END AS validacion_fecha

FROM Results_new;

-----

Select

cnpj,

GETDATE() hoje,

DATEDIFF(year, data_abertura_empresa, getdate()) idade

from Results

-----

```

## Conclusão

Foi uma jornada complicada pela falta de familiaridade com as ferramentas da Amazon, porém ao final do trabalho de ETL e análise, apesar de muito simples, pude aprender a lógica e possibilidades que nos trazem as aplicações de ETL e os ambientes de serviço de análise por queries SQL. A oportunidade que tive de fazer um trabalho prático me abriu os olhos para as possibilidades para o futuro.