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

PREFEITURA 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.

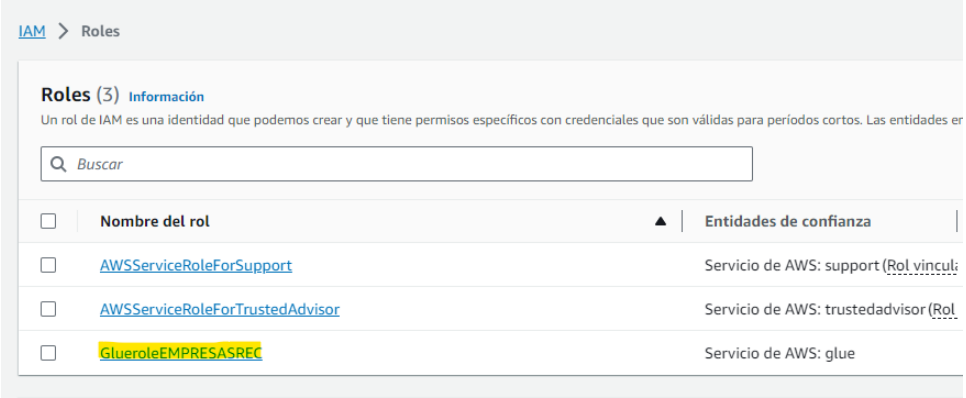
Ferramentas utilizadas para ETL e Queries

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.

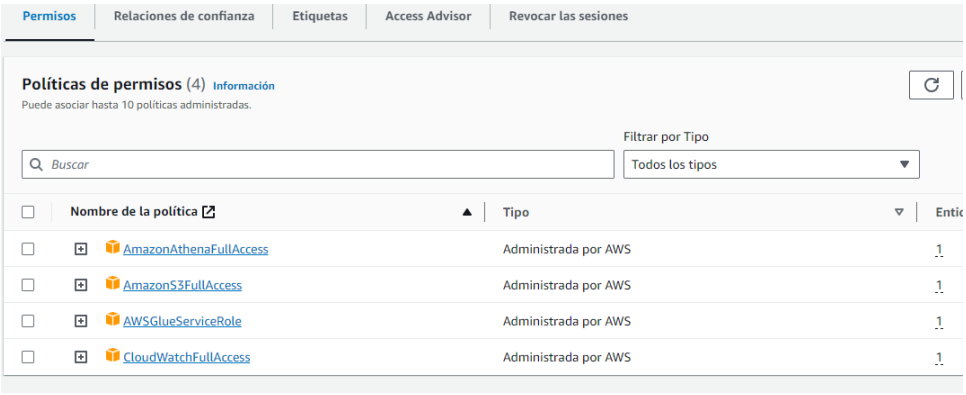
Passo a passo do projeto

1 – Criação de Role e Configuração das permissões:

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



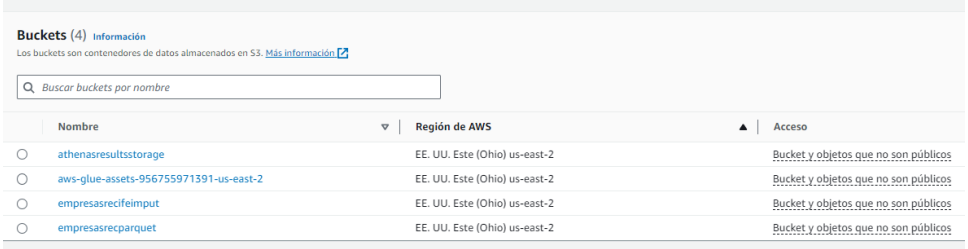
Em seguida configurei as permissoes necessárias para utilizar S3, AWS Glue e Athena:



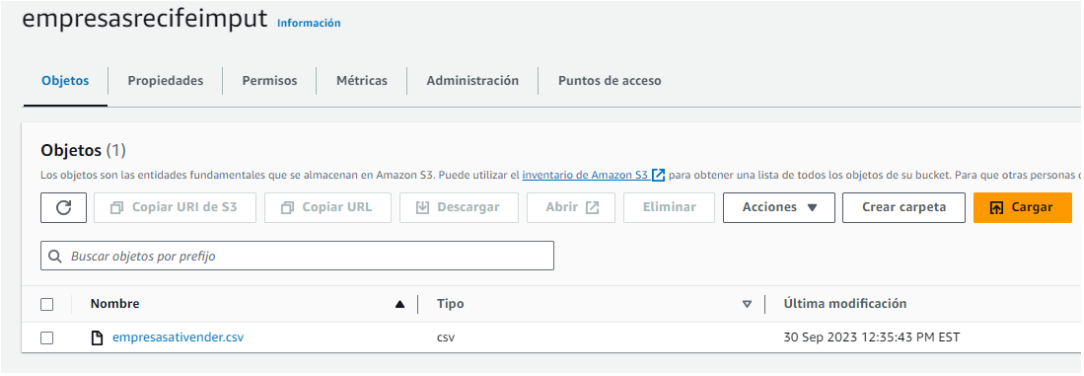
2 – Criacao e alimentacao de Buckets in S3

Criei 3 buckets:

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



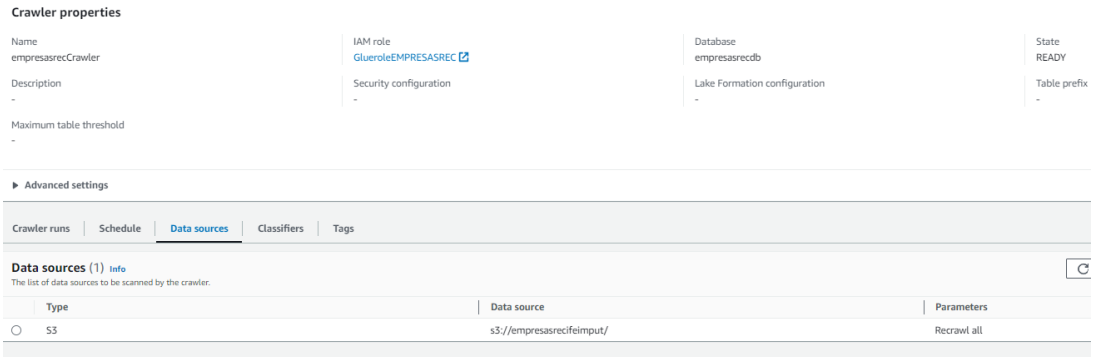
3 – Upload de arquivo empresasativender.csv ao bucket empresasrecimput:



4 – Criacao de Crawler para arquivo .csv:

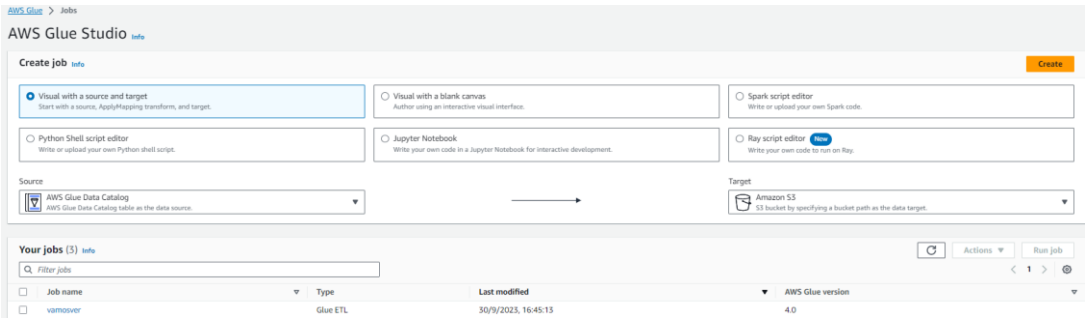
A criação do crawler é necessária para inspeção do arquivo como um todo, para classificar, agrupar dados em tabelas ou partições e gravar metadados no Data Catalog.

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



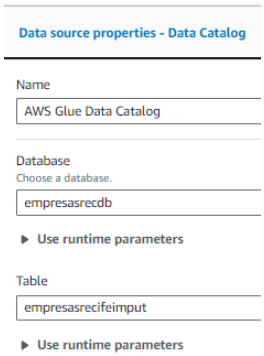
5 – Criacao de ETL JOB:

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

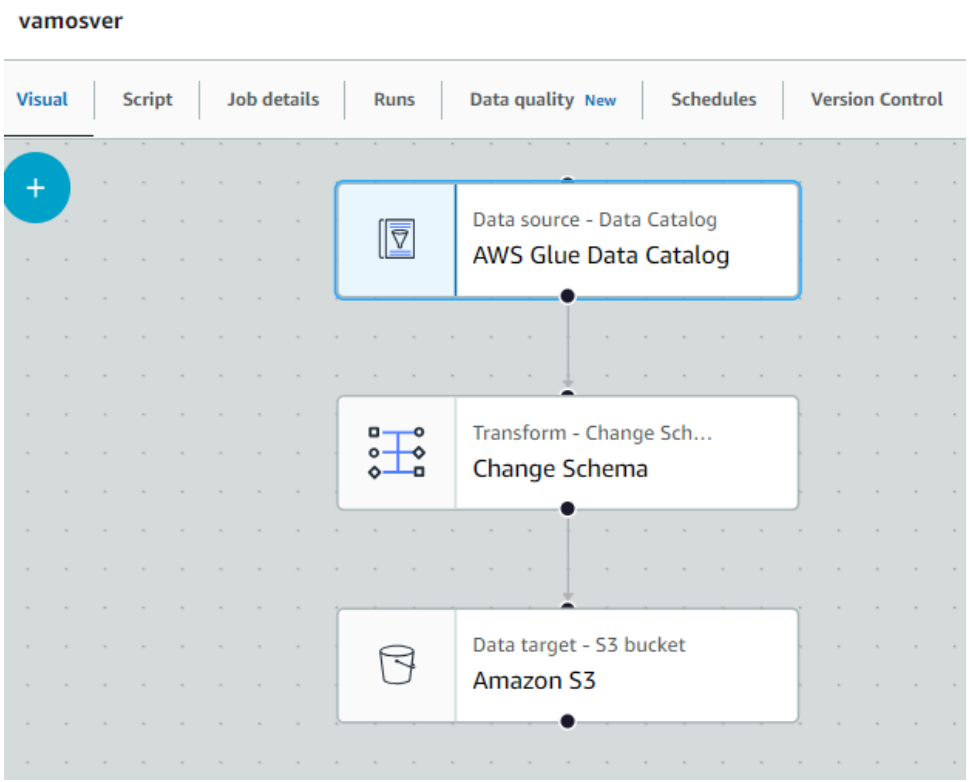


6 – Criação de pipeline visual de ETL:

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



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

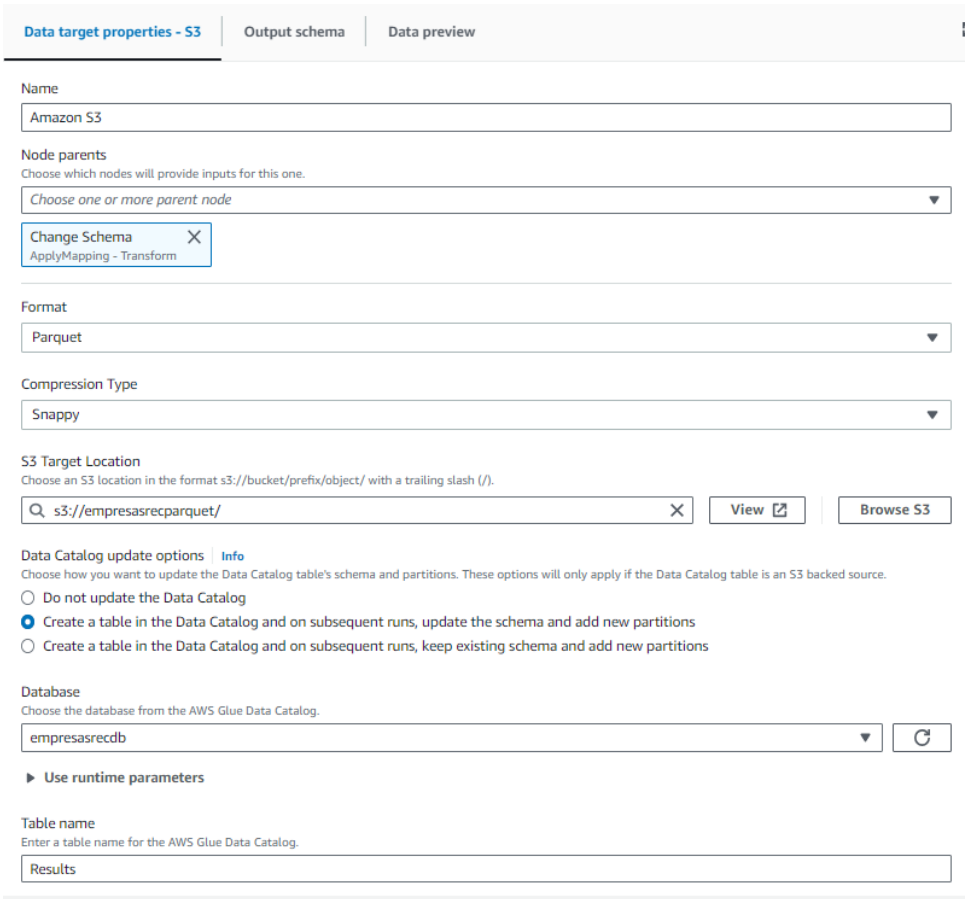


Para logo fazer os “drops” das colunas que me pareciam dispensáveis:

7 – Definicao de output:

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

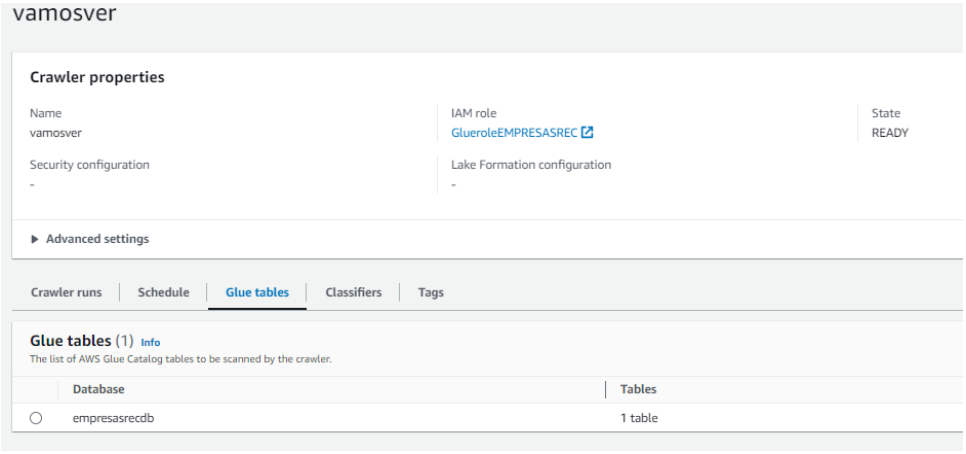
Coloquei como output uma tabela chamada “results”.



8 – Execução de ETL JOB

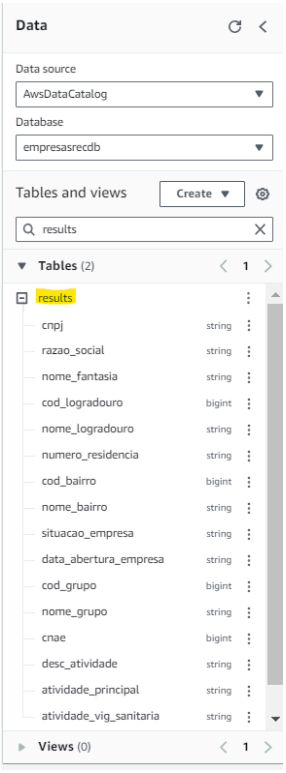
9 – Execução de crawler para DB

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



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.



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 atuacao (nome_grupo) e dentro dessas areas há 1232 atividades economicas (desc_atividades).

Toda empresa tem CNPJ, Razao Social e Nome Fantasia. E 1,17% das empresas possuem Razao 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 adjancente ao mesmo. Em terceiro lugar está Boa Vista (20285 empresas) que é parte do centro do Recife.

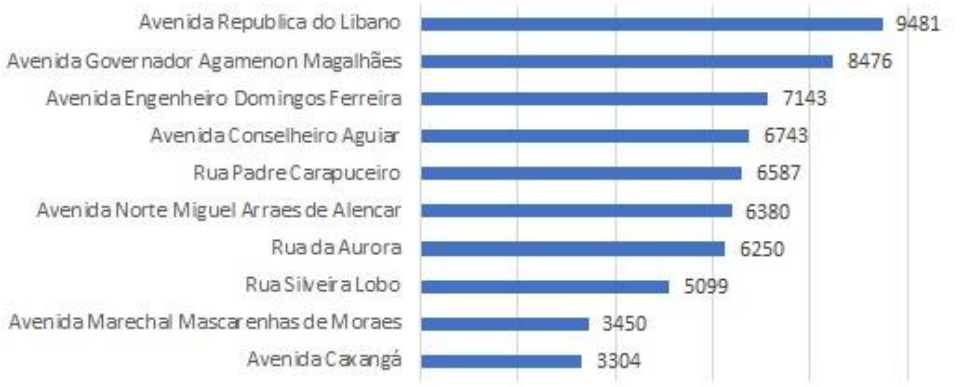
Boa Viagem e Pina:

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 incidencia 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 “servicos 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 atuacao:

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

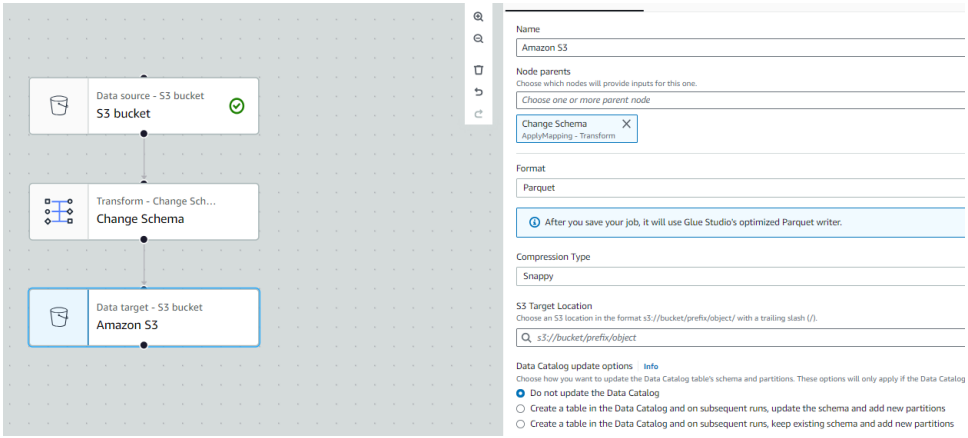
QUERY UTILIZADA:

```
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

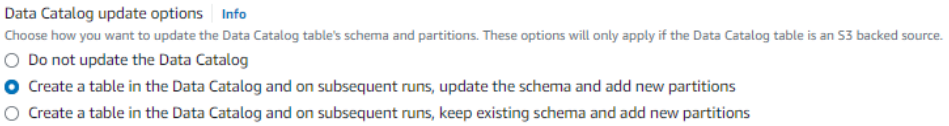
1 - Criação do Pipeline – Source Setting e Arquivo Parquet valores vazios

Durante a fase de criacao 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 opcao de Data Catalog options:



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 – Funcao datediff

Tive a intenção de calcular o tempo 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 aplicacoes de ETL e os ambientes de servico de análise por queries SQL. A oportunidade que tive de fazer um trabalho prático me abriu os olhos para as possibilidades para o futuro.