

MVP - Sprint III – Engenharia de Dados

Regina Delcourt

Avaliação de Desastres Naturais

Sumário

- Descrição3
- Objetivo.....3
- Detalhamento3
 - 1. Busca pelos dados.....3
 - 2. Coleta4
 - 3. Modelagem.....7
 - 4. Carga e ETL.....12
 - 5. Análise21
- Conclusão29
 - Autoavaliação29

Descrição

Neste trabalho, é apresentada a construção de um pipeline de dados utilizando tecnologia na nuvem da AWS (Amazon Web Services). O pipeline envolve a busca, coleta, modelagem, carga e análise dos dados.

Objetivo

O objetivo do presente trabalho é a avaliação da ocorrência de desastres naturais no Brasil.

A partir dessa avaliação, pretende-se responder às seguintes perguntas:

- 1 - Qual foi o tipo de evento com maior número de óbitos?
- 2 - Qual foi o tipo de evento com maior prejuízo financeiro?
- 3 - Qual foi o tipo de evento com maiores danos humanos?
- 4 - Esses eventos coincidem?
- 5 - Quais os municípios com maior ocorrência desse tipo de evento?
- 6 - Esses municípios pertencem ao mesmo estado brasileiro?
- 7 - Caso contrário, qual estado brasileiro possui predominância desses eventos?
- 8 - É possível prever a qual fator natural esse evento pode ser associado?

Detalhamento

1. Busca pelos dados

Buscando avaliar a ocorrência de desastres naturais no Brasil foi utilizado o Sistema Integrado de Informações sobre Desastres - S2ID.

O S2ID integra diversos produtos da Secretaria Nacional de Proteção e Defesa Civil – SEDEC, e tem como objetivo qualificar e dar transparência à gestão de riscos e desastres no Brasil, por meio da informatização de processos e disponibilização de informações sistematizadas dessa gestão.

Para tal foi acessado o Atlas Digital de Desastres no Brasil, por meio do seguinte endereço eletrônico:

<http://atlasdigital.mdr.gov.br/paginas/downloads.xhtml>



Figura 1 – Busca pelos Dados

2. Coleta

Foram utilizados os dados em formato CSV para facilitar o trabalho com os mesmos na nuvem, além de, conforme informado no site do Atlas Digital de Desastres no Brasil, os dados nesse formato já considerarem a conversão dos valores monetários corrigidos para dezembro de 2022.

Os dados foram baixados para a máquina local e foram divididos em dois arquivos, de forma a possibilitar mais opções de transformação na nuvem (ex. Join de arquivos).

Após a realização de teste iniciais das três nuvens sugeridas para realização do trabalho, a saber, AWS (Amazon Web Services), Microsoft Azure e Google Cloud Storage (GCS), decidiu-se por utilizar os serviços da AWS.

Após alguns testes iniciais foi constatada a necessidade de transformar o arquivo CSV. O arquivo original possuía delimitador por “;”, e as plataformas de nuvem apresentavam erro na leitura do arquivo. Dessa forma, conforme apresentado a seguir, foi feita a transformação do arquivo por meio do Google Colab, utilizando Python, para arquivo CSV com delimitador por ‘,’.

```
import csv

# Nome do arquivo de entrada e de saída
arquivo_entrada = 'BD_Atlas_1991_2022_1_2.csv'
arquivo_saida = 'BD_Atlas_1991_2022_1_2a.csv'

# Abra o arquivo de entrada e o arquivo de saída
with open(arquivo_entrada, 'r', newline='', encoding='utf-8') as entrada, \
    open(arquivo_saida, 'w', newline='', encoding='utf-8') as saida:

    # Crie objetos CSV para entrada e saída
    leitor_csv = csv.reader(entrada, delimiter=';')
    escritor_csv = csv.writer(saida, delimiter=',')

    # Copie os dados do arquivo de entrada para o arquivo de saída
    for linha in leitor_csv:
        escritor_csv.writerow(linha)

print(f'O arquivo {arquivo_saida} foi criado com sucesso!')
```

O arquivo BD_Atlas_1991_2022_1_2a.csv foi criado com sucesso!

Figura 2 – Transformação do primeiro arquivo utilizando Python no Google Colab

```
import csv

# Nome do arquivo de entrada e de saída
arquivo_entrada = 'BD_Atlas_1991_2022_2_2.csv'
arquivo_saida = 'BD_Atlas_1991_2022_2_2a.csv'

# Abra o arquivo de entrada e o arquivo de saída
with open(arquivo_entrada, 'r', newline='', encoding='utf-8') as entrada, \
    open(arquivo_saida, 'w', newline='', encoding='utf-8') as saida:

    # Crie objetos CSV para entrada e saída
    leitor_csv = csv.reader(entrada, delimiter=';')
    escritor_csv = csv.writer(saida, delimiter=',')

    # Copie os dados do arquivo de entrada para o arquivo de saída
    for linha in leitor_csv:
        escritor_csv.writerow(linha)

print(f'O arquivo {arquivo_saida} foi criado com sucesso!')
```

O arquivo BD_Atlas_1991_2022_2_2a.csv foi criado com sucesso!

Figura 3 – Transformação do segundo arquivo utilizando Python no Google Colab

Após a conversão dos arquivos, deu-se início ao trabalho na nuvem da AWS, com a criação do “bucket” onde o projeto foi trabalho, chamado de “mvpengdados”, conforme Figura 4.

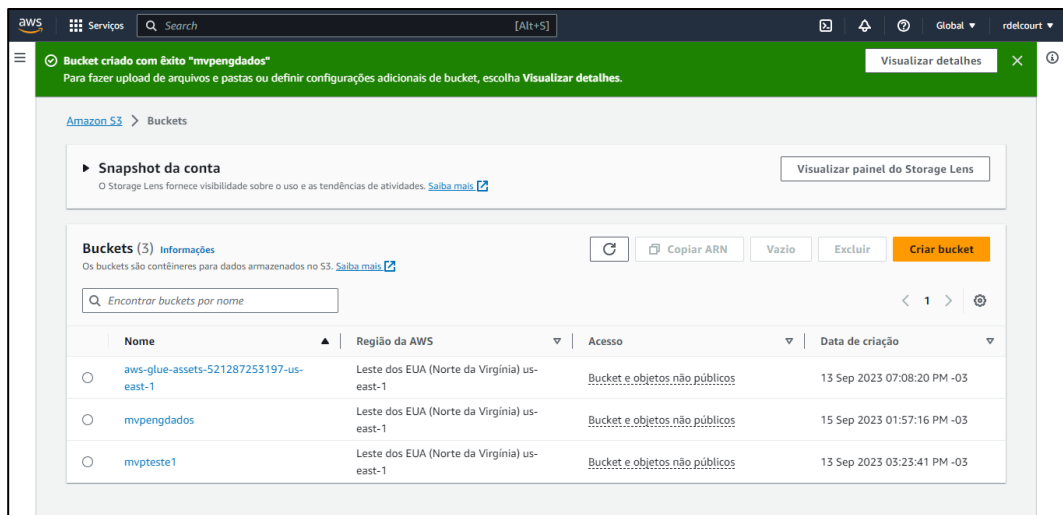


Figura 4 – *Bucket* criado no Amazon S3

Dentro do *bucket* mvpengdados, criou-se a pasta de objetos (pasta de trabalho) “mvpengdados”, onde o presente trabalho foi realizado, conforme figura a seguir.

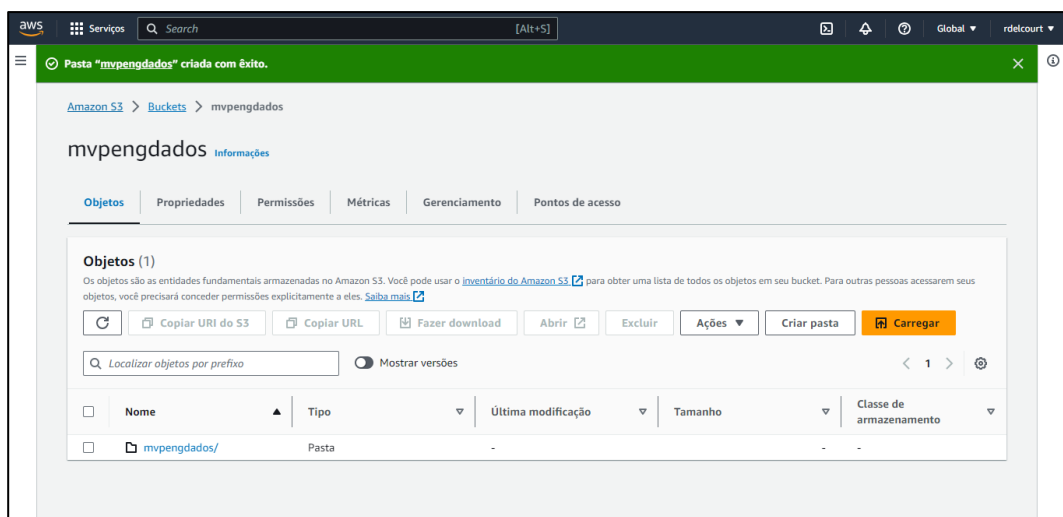


Figura 5 – Pasta de trabalho no *Bucket* do Amazon S3

Os dados foram, então, inseridos e armazenados manualmente no *bucket* criado, conforme Figura 6.

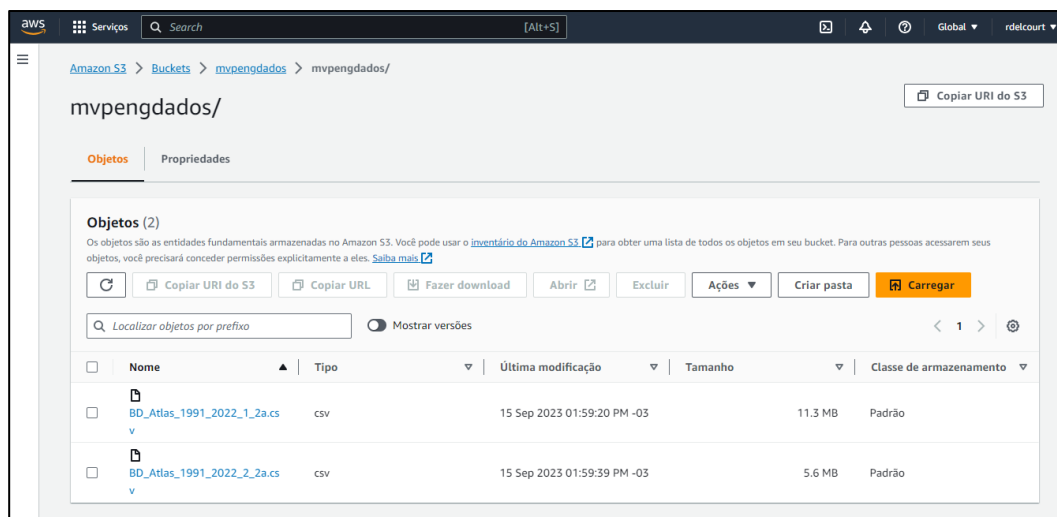


Figura 6 – Objetos na pasta de trabalho no *Bucket* do Amazon S3

3. Modelagem

Para o presente trabalho foi considerado um modelo de dados em formato *flat* (*Data Lake*). Os dados são armazenados em seu formato bruto, *flat*, em arquivos CSV. Os dados são armazenados sem uma estrutura fixa e podem ser processados e transformados conforme necessário.

A Figura 7 apresenta a modelagem de dados utilizada.

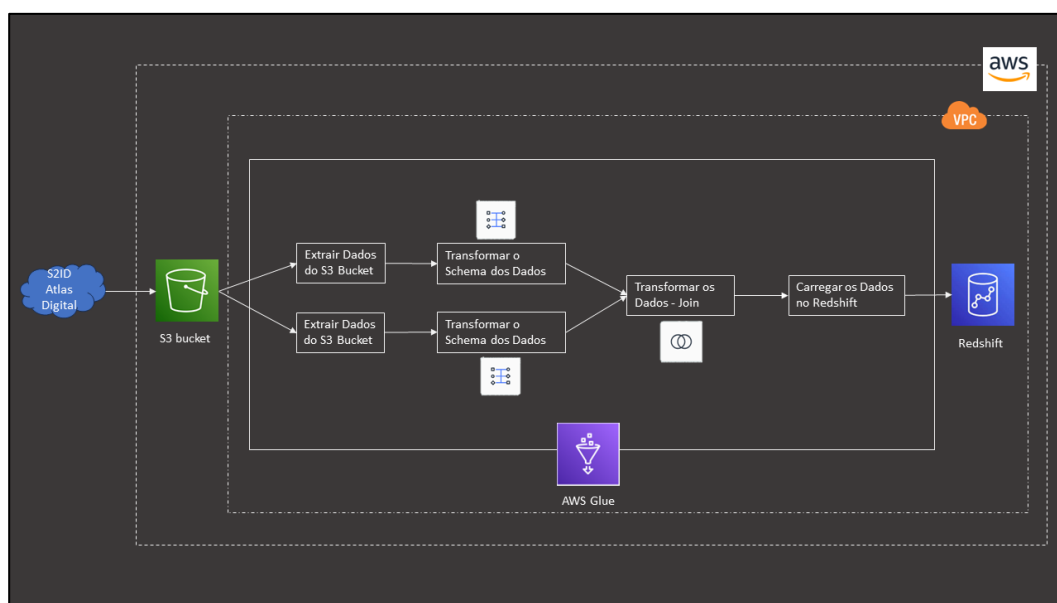


Figura 7 - Modelagem de Dados

No presente trabalho foi utilizada a ferramenta “*Glue*” da AWS, para Extração, Transformação e Carga (ETL), onde os dados podem ser processados e transformados, utilizando ferramentas de “*Big Data*”, como o “*Apache Spark*”, para análise posterior.

Utilizando-se a *AWS Glue* foi construída uma base de dados (Figura 8) com uma tabela (Figura 9) de Catálogo de Dados (Figura 10), contendo uma descrição dos dados.

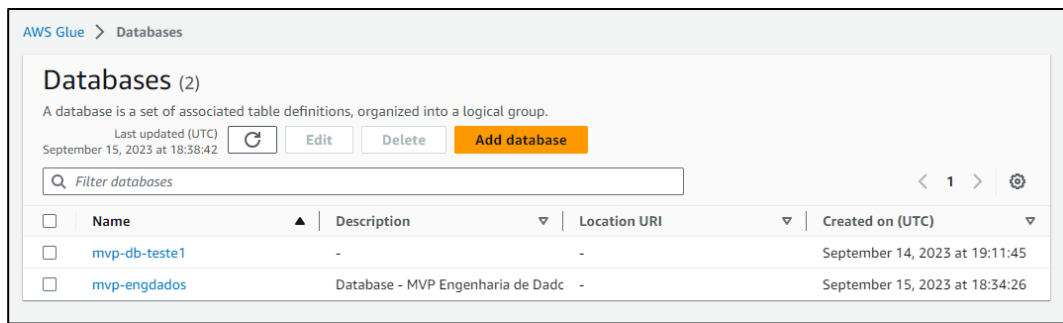


Figura 8 - Base de Dados

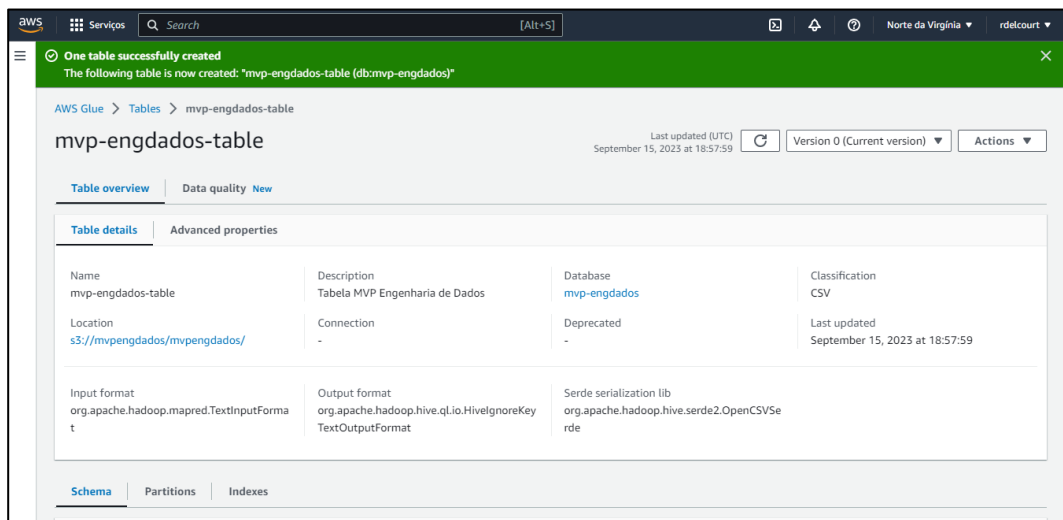


Figura 9 – Detalhes da Tabela de Dados

Foi construído um Catálogo de Dados contendo uma descrição detalhada dos dados e seus domínios, conforme apresentado na figura a seguir.

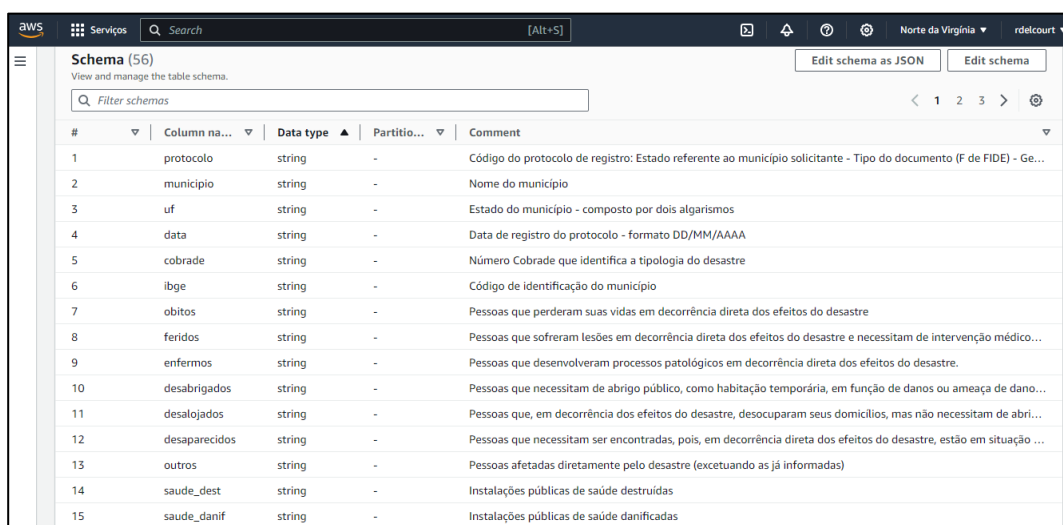


Figura 10 - Catálogo de Dados com descrição

Foi, ainda, testada a função de qualidade de dados do AWS *Glue*, com criação de regras de qualidade de dados.



Figura 11 – Criação de Regras de Qualidade de Dados

A Figura 12 apresenta o Código referente às regras de Qualidade de Dados criado pelo AWS *Glue*.

```
Rules = [
  RowCount between 62274 and 249096,
  IsComplete "protocolo",
  ColumnLength "protocolo" between 8 and 28,
  IsComplete "municipio",
  ColumnLength "municipio" between 2 and 33,
  IsComplete "uf",
  ColumnValues
    "uf"
    in
["MG","RS","SC","BA","PB","PI","PE","CE","RN","PR","SP","ES","MS","MT","RJ","PA","AL","AM","MA","SE","GO",
"TO","AC","RR","RO","AP","DF","pa","uf"],
  ColumnValues "uf" in ["MG","RS","SC","BA","PB","PI","PE","CE","RN","PR","SP","ES","MS","MT","RJ","PA"] with
threshold >= 0.91,
  ColumnLength "uf" = 2,
  IsComplete "data",
  ColumnLength "data" between 3 and 11,
  IsComplete "cobrade",
  ColumnValues
    "cobrade"
    in
["Nordeste","14110","Sul","Sudeste","12200","14120","12100","13214","13215","Centro-oeste"] with threshold
>= 0.9,
  ColumnLength "cobrade" between 2 and 13,
  IsComplete "ibge",
  ColumnLength "ibge" <= 9,
  IsComplete "obitos",
  ColumnValues "obitos" in ["0","Estiagem e Seca","Enxurradas","Inundações","Chuvas Intensas","Vendavais e
Ciclones"] with threshold >= 0.91,
  ColumnLength "obitos" <= 31,
  IsComplete "feridos",
  ColumnLength "feridos" <= 17,
```

Completeness "enfermos" >= 0.49,
 ColumnLength "enfermos" <= 8,
 Completeness "desabrigados" >= 0.49,
 ColumnLength "desabrigados" <= 12,
 Completeness "desalojados" >= 0.49,
 ColumnLength "desalojados" <= 11,
 Completeness "desaparecidos" >= 0.49,
 ColumnLength "desaparecidos" <= 13,
 Completeness "outros" >= 0.49,
 ColumnLength "outros" <= 7,
 Completeness "saude_dest" >= 0.49,
 ColumnValues "saude_dest" in ["0"] with threshold >= 0.99,
 ColumnLength "saude_dest" <= 10,
 Completeness "saude_danif" >= 0.49,
 ColumnLength "saude_danif" <= 11,
 Completeness "saude_valor" >= 0.49,
 ColumnLength "saude_valor" <= 11,
 Completeness "ensino_dest" >= 0.49,
 ColumnValues "ensino_dest" in ["0"] with threshold >= 0.99,
 ColumnLength "ensino_dest" <= 11,
 Completeness "ensino_danif" >= 0.49,
 ColumnValues "ensino_danif" in ["0"] with threshold >= 0.95,
 ColumnLength "ensino_danif" <= 12,
 Completeness "ensino_valor" >= 0.49,
 ColumnLength "ensino_valor" <= 12,
 Completeness "outros_dest" >= 0.49,
 ColumnValues "outros_dest" in ["0",""] with threshold >= 0.99,
 ColumnLength "outros_dest" <= 11,
 Completeness "outros_danif" >= 0.49,
 ColumnValues "outros_danif" in ["0",""] with threshold >= 0.99,
 ColumnLength "outros_danif" <= 12,
 Completeness "outros_valor" >= 0.49,
 ColumnLength "outros_valor" <= 12,
 Completeness "comuni_dest" >= 0.49,
 ColumnLength "comuni_dest" <= 11,
 Completeness "comuni_danif" >= 0.49,
 ColumnLength "comuni_danif" <= 12,
 Completeness "comuni_valor" >= 0.49,
 ColumnLength "comuni_valor" <= 12,
 Completeness "hab_dest" >= 0.49,
 ColumnLength "hab_dest" <= 8,

Completeness "habt_danif" >= 0.49,
ColumnLength "habt_danif" <= 10,
Completeness "hab_valor" >= 0.49,
ColumnLength "hab_valor" <= 11,
Completeness "infra_dest" >= 0.49,
ColumnLength "infra_dest" <= 10,
Completeness "infra_danif" >= 0.49,
ColumnLength "infra_danif" <= 11,
Completeness "infra_valor" >= 0.49,
ColumnLength "infra_valor" <= 11,
Completeness "agricultura" >= 0.49,
ColumnLength "agricultura" <= 11,
Completeness "pecuaria" >= 0.49,
ColumnLength "pecuaria" <= 11,
Completeness "industria" >= 0.49,
ColumnLength "industria" <= 11,
Completeness "servicos" >= 0.49,
ColumnLength "servicos" <= 11,
Completeness "total_privado" >= 0.49,
ColumnLength "total_privado" <= 13,
Completeness "saude" >= 0.49,
ColumnLength "saude" <= 11,
Completeness "agua" >= 0.49,
ColumnLength "agua" <= 11,
Completeness "esgoto" >= 0.49,
ColumnLength "esgoto" <= 11,
Completeness "limpeza" >= 0.49,
ColumnLength "limpeza" <= 11,
Completeness "pragas" >= 0.49,
ColumnLength "pragas" <= 10,
Completeness "energia" >= 0.49,
ColumnLength "energia" <= 11,
Completeness "telecom" >= 0.49,
ColumnLength "telecom" <= 10,
Completeness "transportes" >= 0.49,
ColumnLength "transportes" <= 11,
Completeness "combustiveis" >= 0.49,
ColumnLength "combustiveis" <= 12,
Completeness "seguranca" >= 0.49,
ColumnLength "seguranca" <= 10,
Completeness "ensino" >= 0.49,

```

ColumnLength "ensino" <= 11,
Completeness "total_publico" >= 0.49,
ColumnLength "total_publico" <= 13,
Completeness "total_danos_materiais" >= 0.49,
ColumnLength "total_danos_materiais" <= 21,
Completeness "total_danos_humanos" >= 0.49,
ColumnValues "total_danos_humanos" in ["false"] with threshold >= 0.99,
ColumnLength "total_danos_humanos" <= 9,
Completeness "prejuizos_totais" >= 0.49,
ColumnLength "prejuizos_totais" <= 19,
Completeness "regiao" >= 0.49,
ColumnLength "regiao" <= 16
]

```

Figura 12 – Código das regras de Qualidade de Dados

4. Carga e ETL

A etapa de ETL foi realizada utilizando-se o “AWS Glue Studio”, conforme Figura 13.

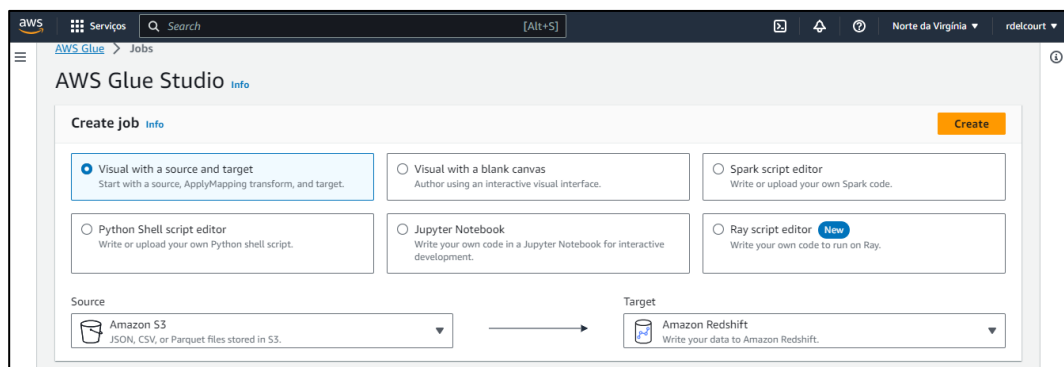


Figura 13 – Criação de Trabalho do “AWS Glue Studio”

Através da interface visual do AWS Glue Studio foram criadas automaticamente as etapas de extração (*source*), transformação (*transform*) e extração (*data target*), conforme apresentado na figura a seguir.

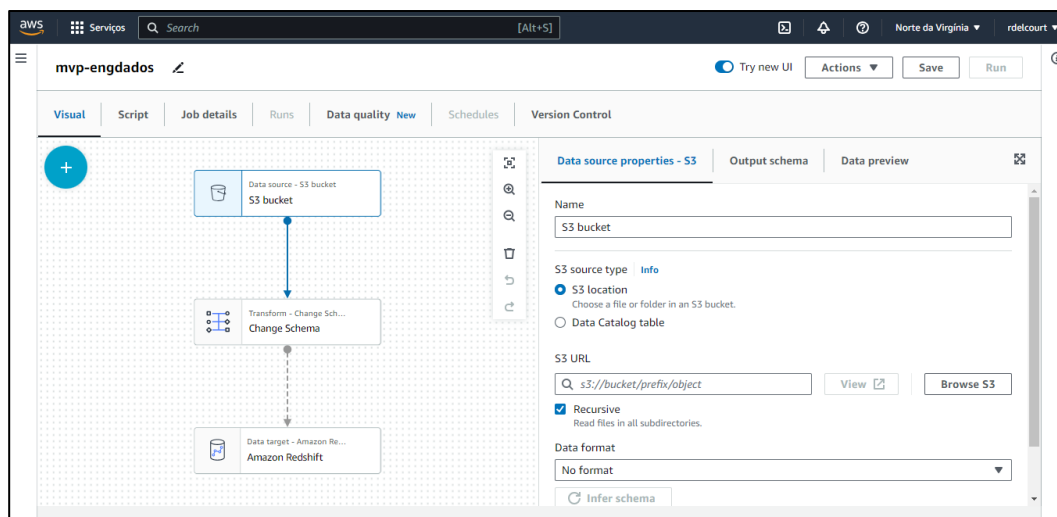


Figura 14 – Interface Visual de Trabalho do AWS Glue inicial

Foi inserida outra fonte de dados, de forma a acrescentar o segundo conjunto de dados, assim como uma etapa adicional de transformação de dados (referente ao segundo conjunto) e a etapa de “join” – para junção e conciliação dos dois conjuntos de dados diferentes. A Figura 15 apresenta a interface visual do trabalho de ETL.

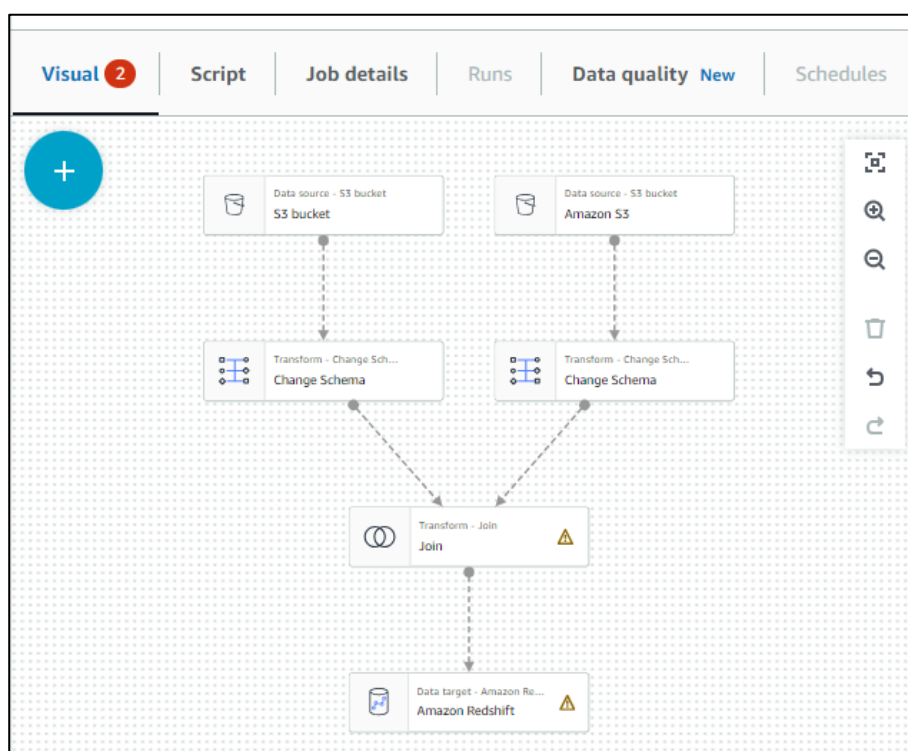


Figura 15 – Modelagem de Trabalho no AWS Glue

Iniciou-se, então, a configuração das etapas ETL, começando-se pelo “Data source – S3 bucket 1_2”, ou seja, foram realizadas as configurações para extração de dados do primeiro conjunto de dados. Foi extraído o arquivo “BD_Atlas_1991_2022_1_2a”, da pasta, “mvpengdados” do bucket “mvpengdados”, conforme figura a seguir.

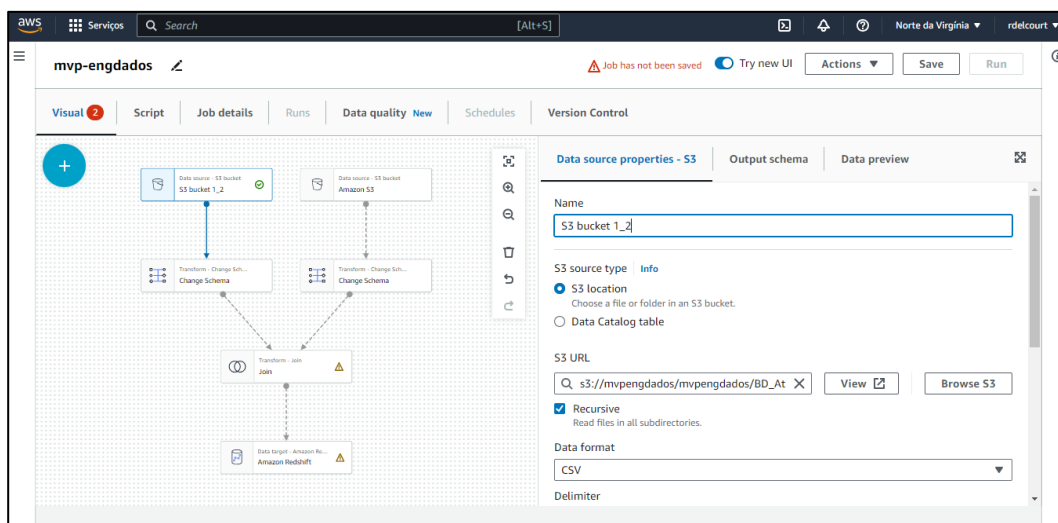


Figura 16 – Configuração do *Data source – S3 bucket 1_2*

A Figura 17 apresenta uma prévia do “*Output do Data source – S3 bucket 1_2*”, ou seja, do esquema de saída do primeiro conjunto de dados extraído no S3 bucket 1_2.

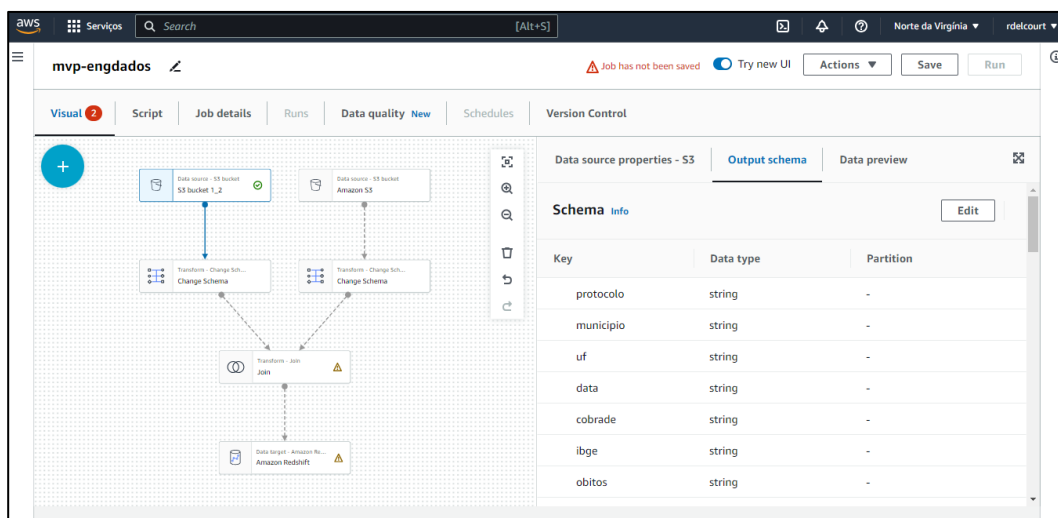


Figura 17 – Prévia do *Output do Data source – S3 bucket 1_2*

De forma semelhante realizou-se a configuração do *Data source – S3 bucket 2_2*. Foram realizadas as configurações para extrair os dados da fonte do segundo conjunto de dados para o S3 bucket 2_2. Foi extraído o arquivo “BD_Atlas_1991_2022_2_2a”, da pasta, “mvpengdados” do bucket “mvpengdados”, conforme figura a seguir.

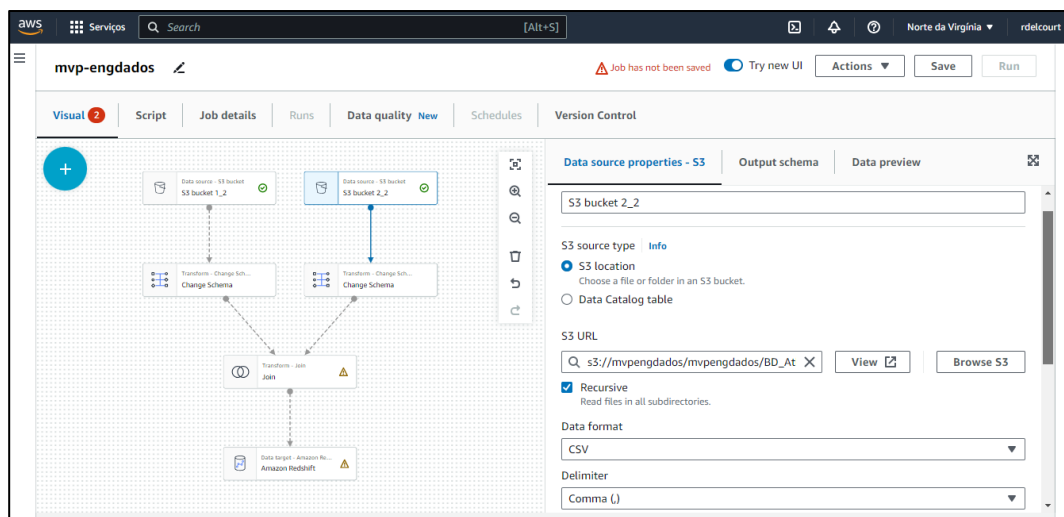


Figura 18 – Configuração do *Data source – S3 bucket 2_2*

A Figura 19 apresenta uma prévia do “*Output do Data source – S3 bucket 2_2*”, ou seja, do esquema de saída do segundo conjunto de dados extraído no *S3 bucket 2_2*.

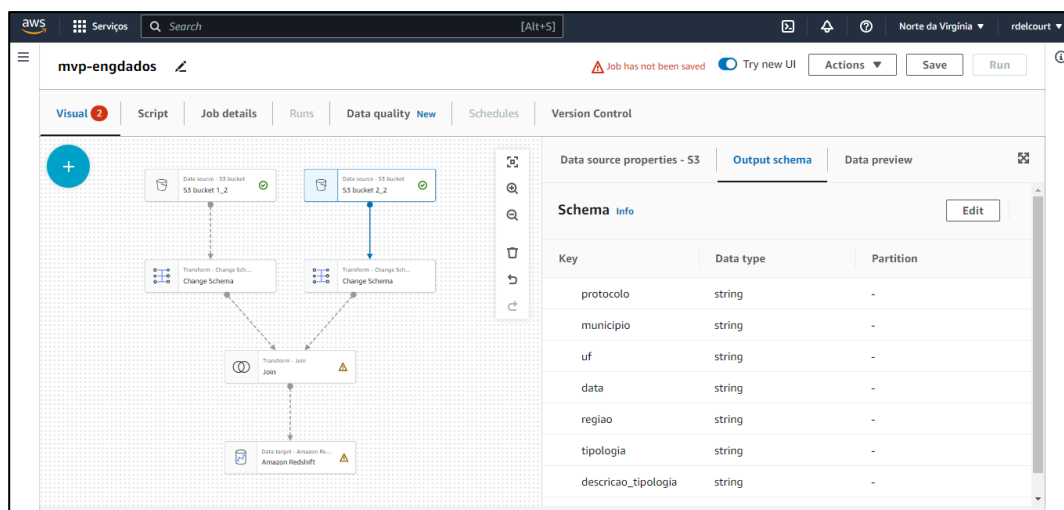


Figura 19 – Prévia do *Output do Data source – S3 bucket 2_2*

Na etapa “*Transform - Change Schema 1_2*”, realizamos a etapa de transformação (*Transform*) dos dados. Foi feita a conversão de alguns tipos dos campos de “*string*” para “*int*”, como, por exemplo, “*óbitos*”, “*feridos*”, “*desabrigados*”, pois os mesmos correspondem ao número de pessoas nas referidas situações. Além da remoção (“*drop*”) de alguns atributos, considerados irrelevantes para o atingimento do objetivo do presente trabalho, como, por exemplo, “*saude_dest*” e “*ensino_dest*”, correspondentes, respectivamente, ao número de instalações de saúde e de ensino destruídas. A Figura 20 apresenta as transformações realizadas para o primeiro conjunto de dados, pertencentes ao *S3 bucket 1_2*.

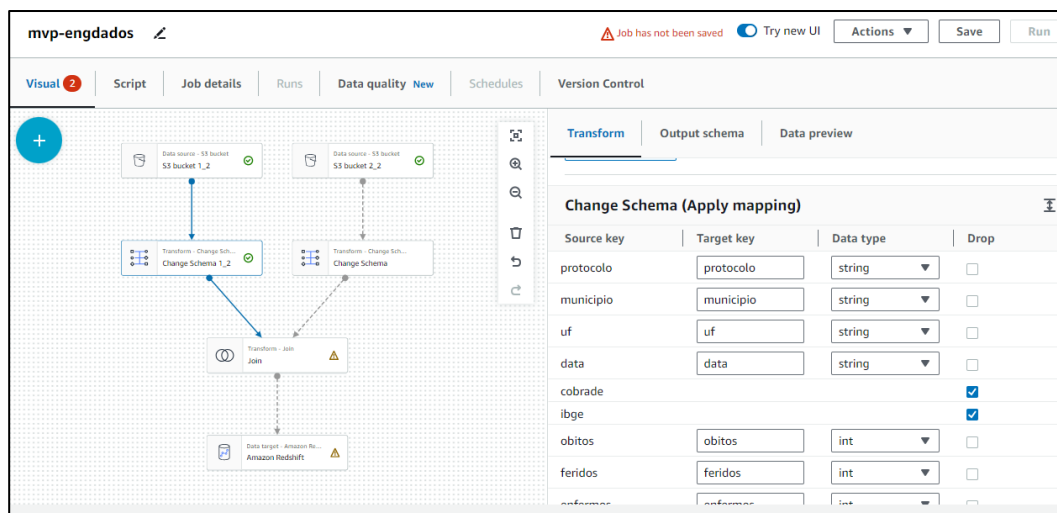


Figura 20 – Etapa de transformação dos dados do S3 bucket 1_2

A Figura 21 apresenta uma prévia "Transform - Change Schema 1_2", ou seja, do esquema de saída da transformação do primeiro conjunto de dados no S3 bucket 1_2.

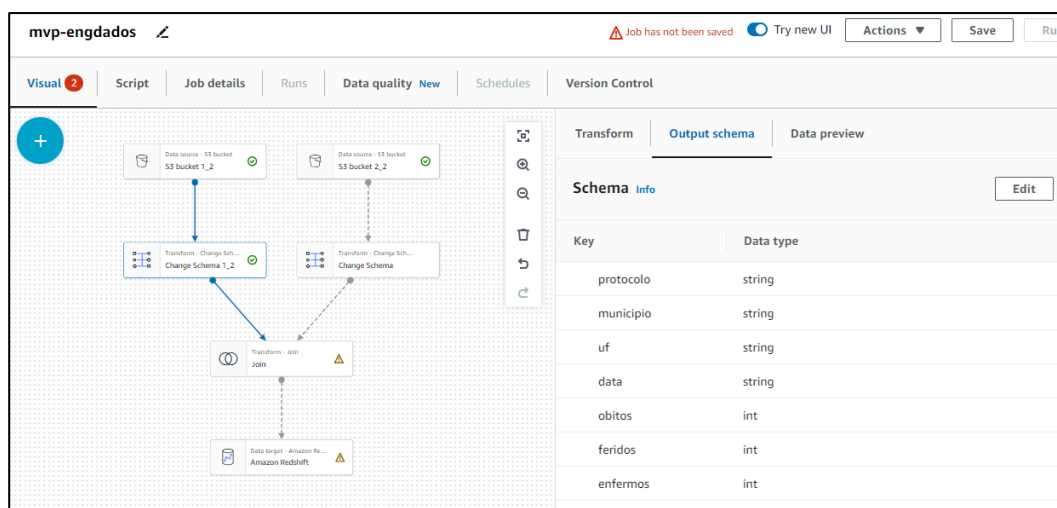


Figura 21 – Prévia do "Transform - Change Schema 1_2"

Na etapa "Transform - Change Schema 2_2", realizamos a etapa de transformação (Transform) dos dados. Foi feita a remoção ("drop") de alguns atributos repetidos dos dados do S3 bucket 1_2 e/ou considerados irrelevantes para o atingimento do objetivo do presente trabalho, como, por exemplo, "uf", "regiao", "tipologia", correspondentes, respectivamente, ao Estado do município, nome da grande região do Brasil e número que identifica a tipologia do desastre. A Figura 22 apresenta as transformações realizadas para o primeiro conjunto de dados, pertencentes ao S3 bucket 1_2.

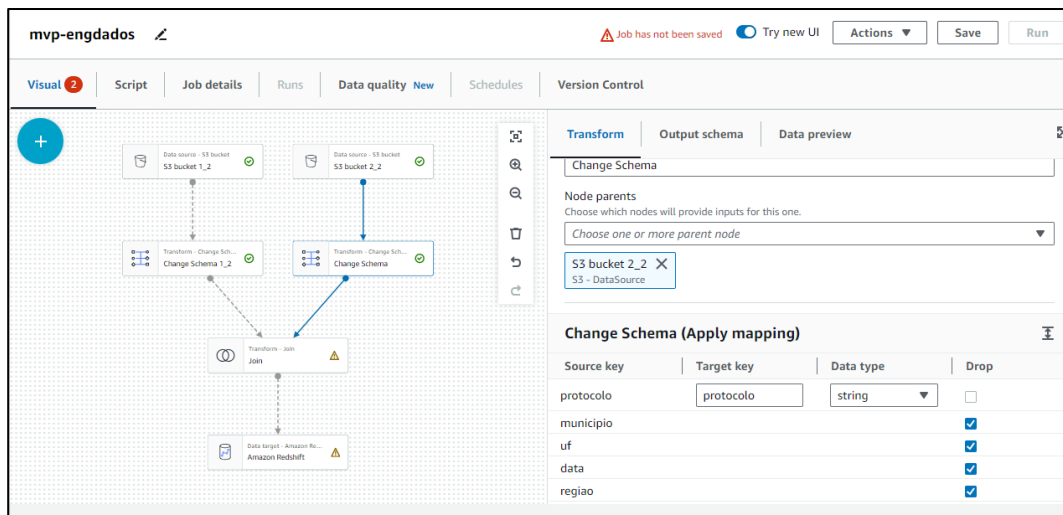


Figura 22 – Etapa de transformação dos dados do S3 *bucket 2_2*

A Figura 23 apresenta uma prévia "Transform - Change Schema 2_2", ou seja, do esquema de saída da transformação do segundo conjunto de dados no S3 *bucket 2_2*.

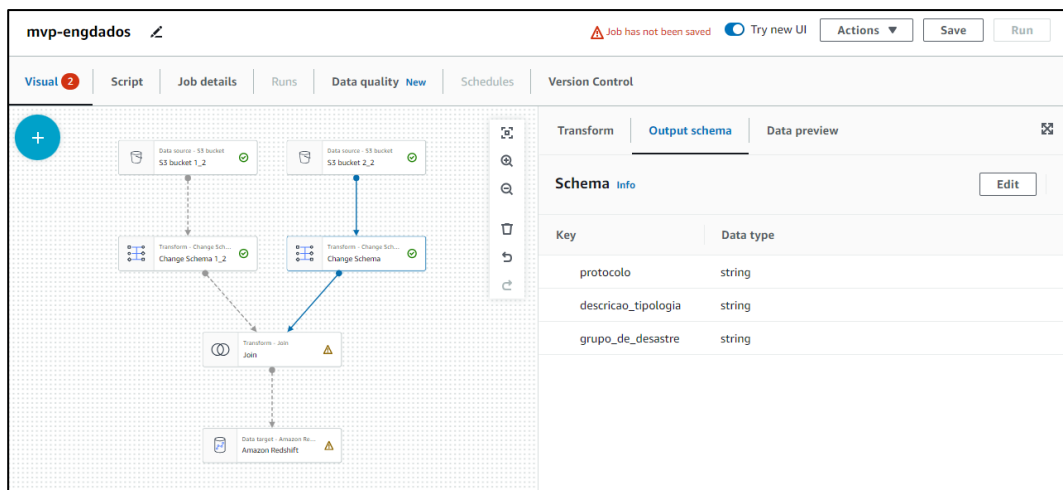


Figura 23 – Prévia do "Transform - Change Schema 2_2"

A ferramenta AWS Glue sugere a inserção de um prefixo nos atributos para evitar a sobreposição dos nomes de cada "parent node", na junção e conciliação dos conjuntos de dados, conforme figura a seguir.

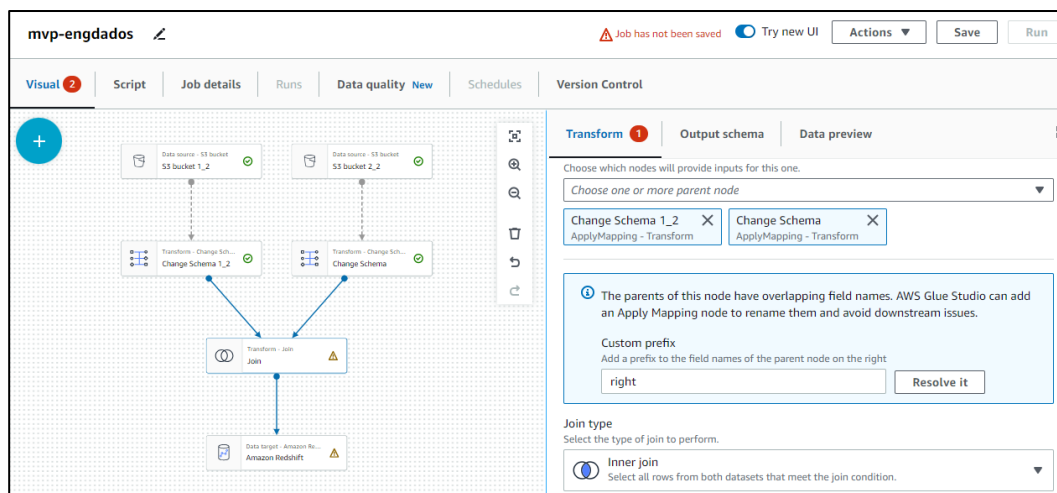


Figura 24 – Inserção de prefixo nos atributos do S3 Bucket 2_2"

Foi selecionado, então, o “parente node” para a junção dos conjuntos de dados, conforme Figura 25.

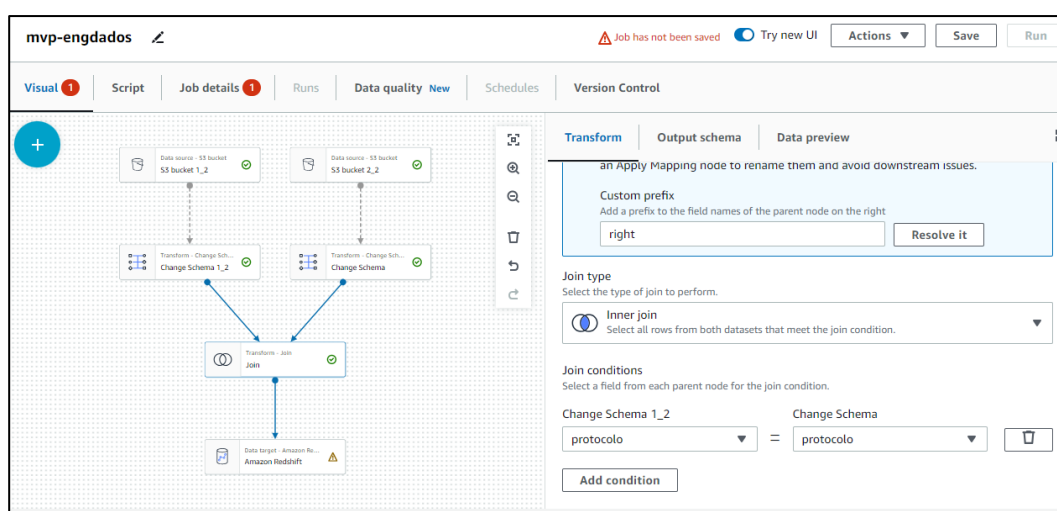


Figura 25 – Seleção do “parente node”

Na etapa de carregamento (Load), "Data target - Amazon Redshift", foram configurados os parâmetros necessários dos dados transformados no nosso banco de dados, conforme figura a seguir.

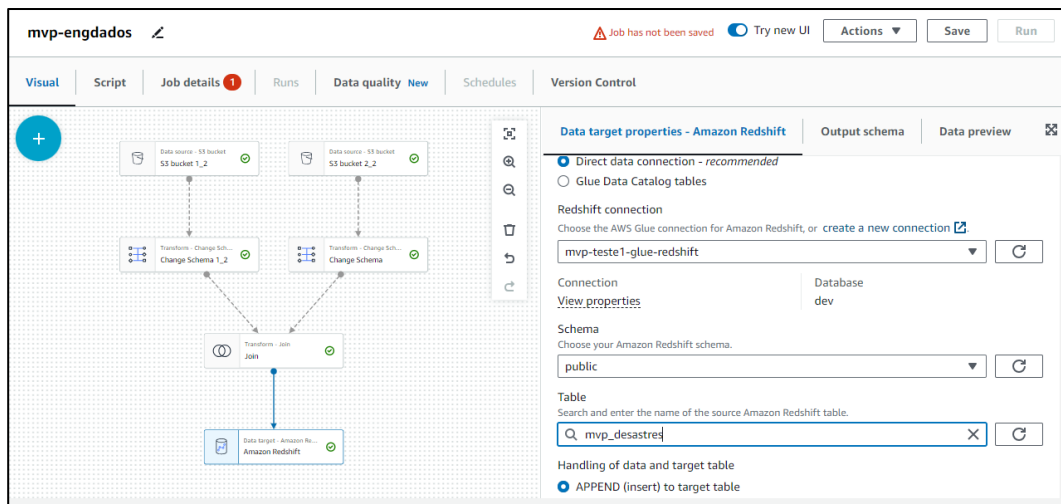


Figura 26 – Configuração do “Data target - Amazon Redshift”

Ainda na etapa de configuração do “Data target - Amazon Redshift” fez-se necessária a criação da tabela fonte do “Amazon Redshift”. A Figura 27 apresenta a criação da referida tabela no “Redshift query editor”.

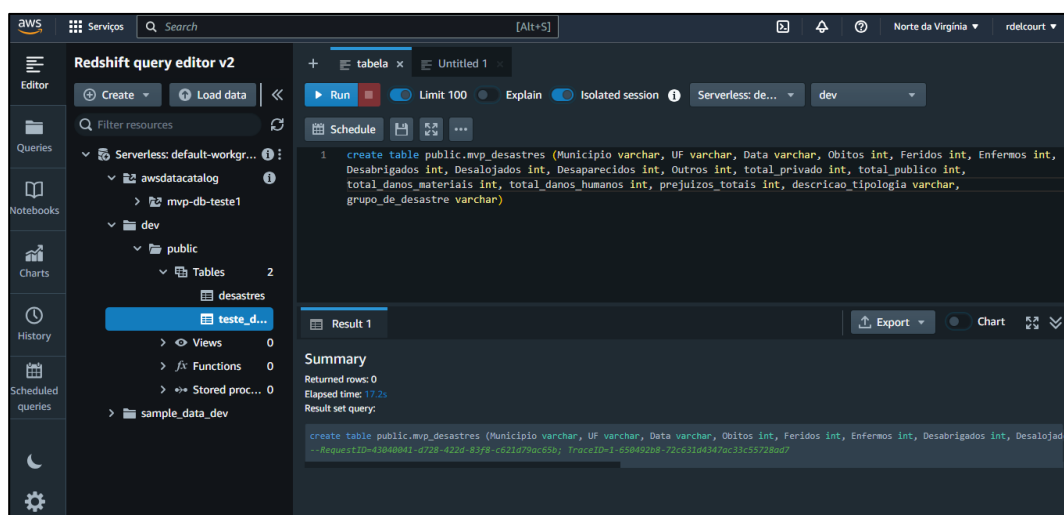


Figura 27 – Criação da tabela fonte do “Data target - Amazon Redshift”

Após as configurações na etapa de carga do “Data target - Amazon Redshift”, os dados foram atualizados e salvos com sucesso, estando prontos para execução, conforme figura a seguir.

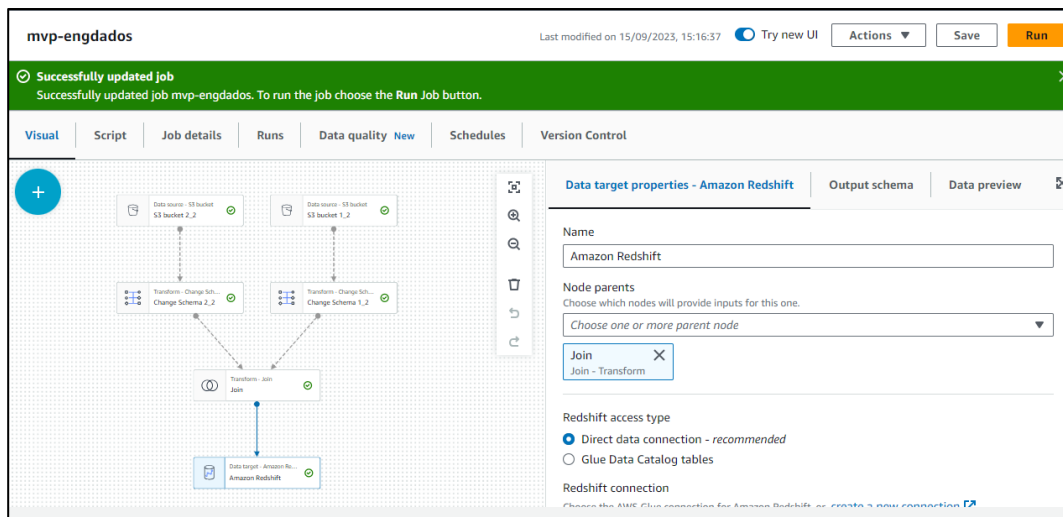


Figura 28 – Dados prontos para execução

O trabalho (“job”) foi, então, executado e registrado, conforme a figura a seguir.

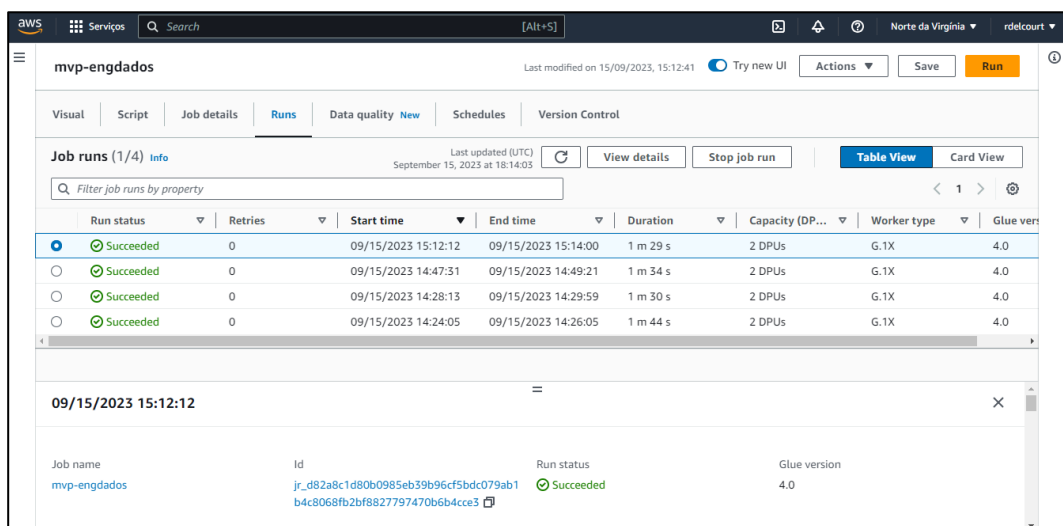


Figura 29 – Execução com sucesso do “job”

A Figura 30 apresenta o “card view” da execução bem-sucedida do trabalho no AWS Glue.

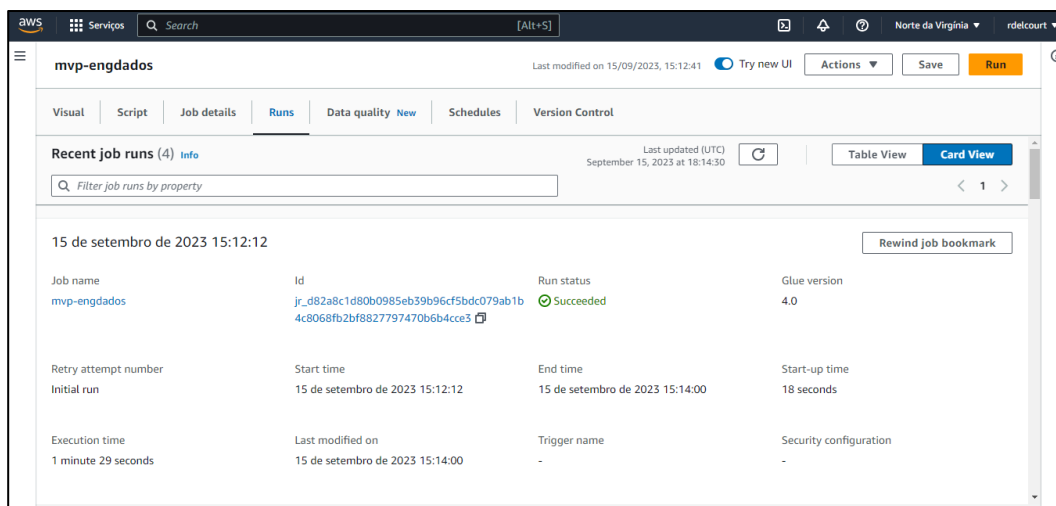


Figura 30 – “Card view” da execução do “job”

5. Análise

A etapa de análise dos dados foi dividida em duas: qualidade de dados e solução do problema.

a. Qualidade de dados

De acordo com o “Manual de Tratamento da Base de Dados do Atlas Digital de Desastres no Brasil”, pertencente ao Sistema Integrado de Informações sobre Desastres (S2ID), disponível no link (<http://atlasdigital.mdr.gov.br/paginas/downloads.xhtml>), o conjunto de dados utilizado no presente trabalho já se encontrava tratado antes de ser disponibilizado.

O Atlas Digital de Desastres no Brasil consiste em um sistema de informação contínua, onde os dados de ocorrências de desastres são historicamente registrados, tratados e estruturados para, somente depois desse processo, serem disponibilizados na plataforma, demandando atualização frequente dos novos registros e a correção de dados já disponíveis.

O Formulário de Informações do Desastre (FIDE), é o documento em que, a partir de 2012, são registradas todas as informações relevantes para a caracterização de um desastre, incluindo a estimativa de afetados, danos materiais e ambientais, e prejuízos públicos e privados. Porém, destaca-se que, independentemente da duração de um desastre, as informações representam, em sua maioria, a realidade até o momento de registro. Isto significa que, se não houver correção dos dados já disponíveis, não há atualização da condição dos afetados (por exemplo, um desaparecido que eventualmente possa ter vindo a óbito), e dos danos materiais e prejuízos resultantes no decorrer do desastre.

O tratamento dos dados, pelo S2ID, foi pautado em três principais etapas: obtenção dos dados, tratamento da base de dados e a atualização dos valores monetários.

Obtenção de dados: os dados de desastres ocorridos entre 1991 e 2012 foram obtidos a partir da digitalização dos protocolos registrados no período. Devido à formalização do registro de desastres no S2ID a partir de 2013, os dados desse período em diante foram obtidos por meio de plataforma específica.

Tratamento da base de dados: o tratamento dos dados é orientado a partir de três etapas: análises de erros de preenchimento, conferência de dados duplicados e análises setoriais e regionais.

O objetivo da etapa de análise de erros de preenchimento é identificar inconsistências na base de dados como valores não inteiros preenchidos em alguma categoria de dano humano, valores extremos em geral (tanto muito alto como muito baixo), e tipologias inadequadas de desastres.

A etapa de conferência de dados duplicados tem como objetivo identificar situações em que uma mesma ocorrência é registrada mais de uma vez. Para identificar registros duplicados, são observados aspectos como o município, a data e a Classificação e Codificação Brasileira de Desastres (Cobrade) da ocorrência.

Um exemplo na conferência de dados duplicados, apontado pelo “Manual de Tratamento da Base de Dados do Atlas Digital de Desastres no Brasil”, foi uma problemática envolvendo os registros das tipologias de seca e estiagem. No Brasil, segundo a Instrução Normativa nº 01, de 24 de agosto de 2012, “O prazo de validade do Decreto que declara a situação anormal decorrente do desastre é de 180 dias a contar de sua publicação em veículo oficial do município ou do estado”. Dessa forma, por se tratar de um desastre gradual, no qual a ocorrência de um único evento pode durar mais de um ano, os registros de uma seca ou estiagem muitas vezes são duplicados. Dessa forma, o S2ID optou por manter os registros que apresentassem maior valor de danos materiais e prejuízos, tendo como resultado, uma redução de 26,3% no número de ocorrências de seca e estiagem registradas, reduzindo de 33.951 registros para 25.024, entre os anos de 1991 e 2019.

O objetivo das análises setoriais e regionais de ocorrências foi identificar as inconsistências relacionadas às tipologias de desastres, principalmente as que têm sua ocorrência condicionadas à alguma região específica, como, por exemplo, a erosão costeira ou ciclones, ambos atrelados ao ambiente costeiro. A partir do georreferenciamento das ocorrências em que foram atribuídas estas tipologias foi possível avaliar a possibilidade de ocorrência ou não dos respectivos desastres em função da posição de cada município no território.

Atualização dos valores monetários: a última etapa de tratamento dos dados consiste na correção monetária dos valores de danos e prejuízos, pois somente com a padronização dos dados monetários é possível fazer a comparação direta entre os valores dos diversos anos existentes na base. Para isso, o Índice Geral de Preços – Disponibilidade Interna (IGP-DI), medido pela Fundação Getúlio Vargas (FGV), foi utilizado. Nele é considerado apenas o período a partir de 1994, após a vigência do Plano Real. Dessa forma, os valores anteriores foram retirados da base de dados (1991 a 1994).

O S2ID ressalta ainda, no “Manual de Tratamento da Base de Dados do Atlas Digital de Desastres no Brasil” que existem limitações tanto de coleta como de tratamento das informações utilizadas. Porém, reconhece a importância e validade dos dados fornecidos, bem como sua aplicabilidade no desenvolvimento de estudos relacionados com as ocorrências e os impactos de desastres no Brasil.

Visto que o conjunto de dados utilizado é curado e bem tratado antes de ser disponibilizado, foi realizado-se uma análise de valores por atributo para verificar a existência de problemas, conforme apresentado na Figura 31.

De forma simplificada analisou-se o número total de registros com o número total de municípios e o número total de registros com tipologia, presentes no conjunto de dados, e verificou-se que os números coincidem, indicando que todos os registros possuem um município e um registro de tipologia de desastre associado.

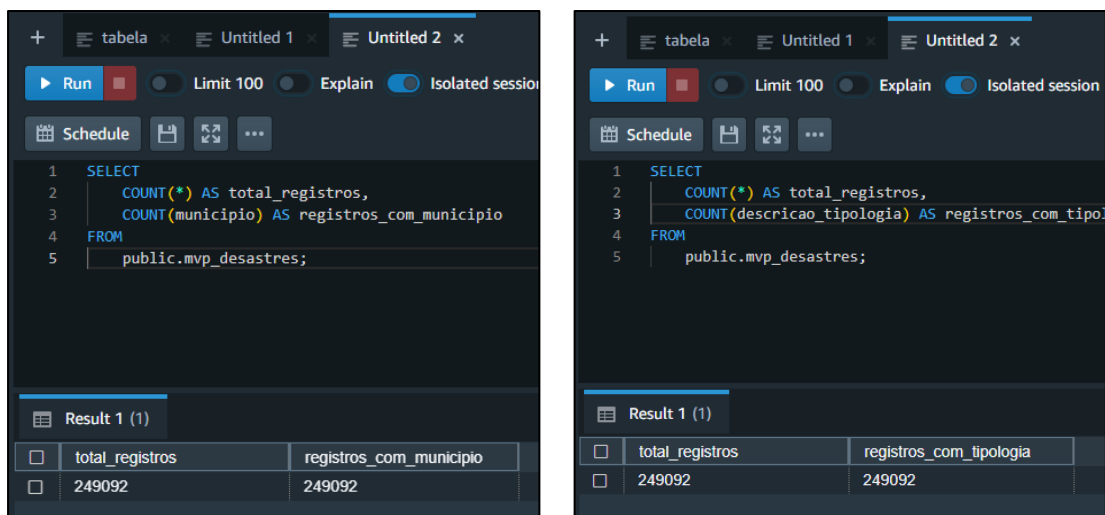


Figura 31 - Análise de valores por atributos

De forma semelhante, conforme figura a seguir, analisou-se o número total de registros com o número de registros com prejuízos, e observou-se uma diferença nos valores. Porém, essa diferença não é um indicativo de má qualidade dos dados, e sim, de que alguns eventos não geraram prejuízos financeiros.

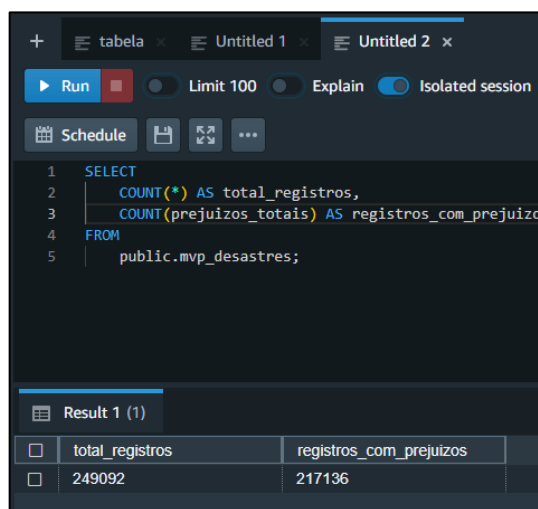


Figura 32 – Análise de número de registros e número de registros com prejuízos

b. Solução do problema

Utilizando-se o “Redshift query editor v2”, da AWS, foi realizada a análise dos dados de forma a responder às perguntas inicialmente apresentadas, como objetivo do presente trabalho.

1 - Qual foi o tipo de evento com maior número de óbitos?

Realizando uma busca em nossa tabela de dados, observa-se que o evento com maior número de óbitos foi a “Enxurrada” ocorrida em janeiro de 2011 no município de Nova Friburgo, no estado do Rio de Janeiro. No evento em questão, classificado no grupo de desastre como um evento do tipo Hidrológico, ocorreram 420 óbitos.

A análise realizada é apresentada na Figura 33.

```
1 SELECT DISTINCT Municipio, UF, data, descricao_tipologia, grupo_de_desastre, obitos
2 from public.mvp_desastres
3 where obitos = (select max(obitos) from public.mvp_desastres)
4
```

Result 1 (1)

municipio	uf	data	descricao_tipologia	grupo_de_desastre	obitos
Nova Friburgo	RJ	13/01/2011	Enxurradas	Hidrológico	420

Figura 33 – Evento com maior número de óbitos

Foi feita, ainda, uma busca para identificar os 10 eventos com maior número de óbitos em nosso banco de dados, conforme Figura 34. Como veremos, novamente, mais adiante, observa-se que a maioria dos eventos com mais fatalidades ocorreram no estado do Rio de Janeiro.

```
1 SELECT DISTINCT Municipio, UF, data, descricao_tipologia, grupo_de_desastre, obitos
2 from public.mvp_desastres
3 ORDER BY obitos desc
4 LIMIT 10;
5
```

Result 1 (10)

municipio	uf	data	descricao_tipologia	grupo_de_desastre	obitos
Nova Friburgo	RJ	13/01/2011	Enxurradas	Hidrológico	420
Teresópolis	RJ	12/01/2011	Enxurradas	Hidrológico	355
Anajás	PA	14/09/1999	Doenças infecciosas	Outros	100
Petrópolis	RJ	15/02/2022	Chuvas Intensas	Hidrológico	78
Petrópolis	RJ	12/01/2011	Enxurradas	Hidrológico	68
Milagres	BA	26/04/2013	Estiagem e Seca	Climatológico	64
Jaboatão dos Guararapes	PE	29/05/2022	Chuvas Intensas	Hidrológico	64
Rio de Janeiro	RJ	05/04/2010	Movimento de Massa	Hidrológico	57
Angra dos Reis	RJ	06/01/2010	Movimento de Massa	Hidrológico	52
Niterói	RJ	05/04/2010	Movimento de Massa	Hidrológico	48

Figura 34 – 10 eventos com maior número de óbitos

2 - Qual foi o tipo de evento com maior prejuízo financeiro?

Primeiramente foi feita uma análise numérica dos prejuízos totais, que incluem tanto os prejuízos públicos quanto os privados. Observa-se, conforme Figura 35, uma média de R\$ 2.548.929 em prejuízos, e um valor máximo de R\$ 1.800.832.000.


```

1 SELECT
2   AVG(prejuizos_totais) AS media_prejuizos,
3   MIN(prejuizos_totais) AS prejuizo_minimo,
4   MAX(prejuizos_totais) AS prejuizo_maximo
5 FROM
6   public.mvp_desastres;

```

Result 1 (1)

	media_prejuizos	prejuizo_minimo	prejuizo_maximo
1	2548929	0	1800832000

Figura 35 – Média e máximo de prejuízo privado e público

Realizando uma busca em nossa tabela de dados, observa-se que o evento com maior prejuízo total, que considera tanto o prejuízo público quanto o privado, foi um evento do grupo de desastre Meteorológico. O evento em questão foram Vendavais e Ciclone, ocorridos em Ponta Porã, no estado do Mato Grosso do Sul, em outubro de 2021.

```

1 SELECT DISTINCT Municipio, UF, data, descricao_tipologia, grupo_de_desastre, prejuizos_totais
2 from public.mvp_desastres
3 where prejuizos_totais = (select max(prejuizos_totais) from public.mvp_desastres)
4

```

Result 1 (1)

municipio	uf	data	descricao_tipologia	grupo_de_desastre	prejuizos_totais
Ponta Porã	MS	26/10/2021	Vendavais e Ciclones	Meteorológico	1800832000

Figura 36 – Evento de maior prejuízo privado e público

Analisando-se os 10 eventos com maior prejuízo total em nosso banco de dados, observa-se uma concentração dos mesmos nos estados da região Centro-Oeste do Brasil. Porém, observa-se, na Figura 37, uma variedade na tipologia e no grupo de desastre desses eventos.

```

1 SELECT DISTINCT Municipio, UF, data, descricao_tipologia, grupo_de_desastre, prejuizos_totais
2 from public.mvp_desastres
3 WHERE prejuizos_totais IS NOT NULL
4 ORDER BY prejuizos_totais desc
5 LIMIT 10;
6

```

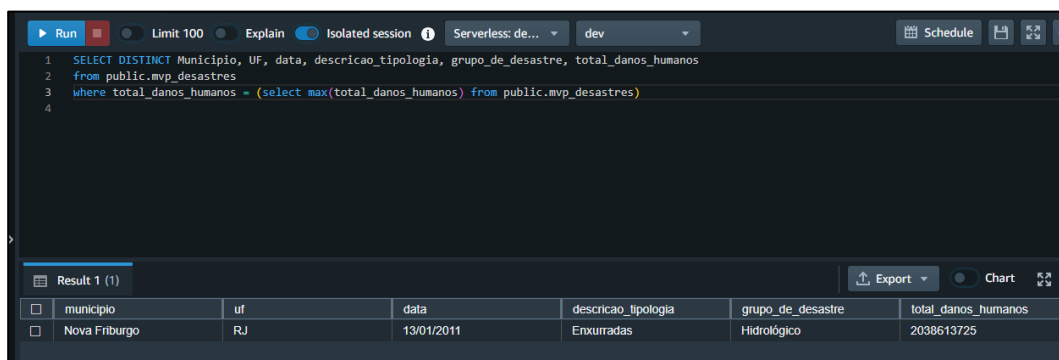
Result 1 (10)

municipio	uf	data	descricao_tipologia	grupo_de_desastre	prejuizos_totais
Ponta Porã	MS	26/10/2021	Vendavais e Ciclones	Meteorológico	1800832000
Dourados	MS	03/01/2022	Estiagem e Seca	Climatológico	1742896767
Maracaju	MS	03/01/2022	Estiagem e Seca	Climatológico	1503663751
Sorriso	MT	11/03/2021	Chuvas Intensas	Hidrológico	1496400000
Nova Ubiratã	MT	22/03/2021	Enxurradas	Hidrológico	1280220764
Boa Vista do Tupim	BA	02/02/2021	Estiagem e Seca	Climatológico	950225000
Ataléia	MG	18/10/2022	Estiagem e Seca	Climatológico	883739250
São Mateus	ES	30/03/2017	Estiagem e Seca	Climatológico	827116049
Ipiranga do Norte	MT	23/03/2021	Chuvas Intensas	Hidrológico	615730000
Corumbá	MS	08/06/2005	Estiagem e Seca	Climatológico	578123000

Figura 37 – 10 eventos de maior prejuízo privado e público

3 - Qual foi o tipo de evento com maiores danos humanos?

Realizando uma busca em nossa tabela de dados, conforme Figura 38, observa-se que o evento com maiores danos humanos corresponde ao evento com maior número de óbitos, sendo este a “Enxurrada” ocorrida em janeiro de 2011 no município de Nova Friburgo, no estado do Rio de Janeiro. No evento em questão, o total de danos humanos remontou a R\$ 20.386.513.725.

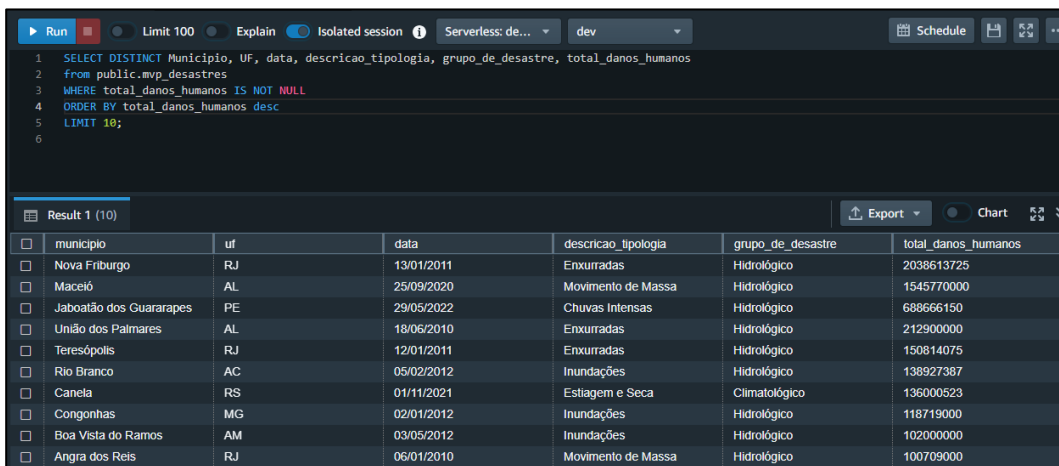


```
1 SELECT DISTINCT Municipio, UF, data, descricao_tipologia, grupo_de_desastre, total_danos_humanos
2 from public.mvp_desastres
3 where total_danos_humanos = (select max(total_danos_humanos) from public.mvp_desastres)
4
```

municipio	uf	data	descricao_tipologia	grupo_de_desastre	total_danos_humanos
Nova Friburgo	RJ	13/01/2011	Enxurradas	Hidrológico	2038613725

Figura 38 – Evento de maior total de danos humanos

Analisando-se, da mesma forma, os 10 eventos com os maiores danos humanos total em nosso banco de dados, apresentado na Figura 39, observa-se quase que uma totalidade de eventos associados ao grupo de desastres Hidrológico.



```
1 SELECT DISTINCT Municipio, UF, data, descricao_tipologia, grupo_de_desastre, total_danos_humanos
2 from public.mvp_desastres
3 WHERE total_danos_humanos IS NOT NULL
4 ORDER BY total_danos_humanos desc
5 LIMIT 10;
6
```

municipio	uf	data	descricao_tipologia	grupo_de_desastre	total_danos_humanos
Nova Friburgo	RJ	13/01/2011	Enxurradas	Hidrológico	2038613725
Maceió	AL	25/09/2020	Movimento de Massa	Hidrológico	1545770000
Jaboatão dos Guararapes	PE	29/05/2022	Chuvas Intensas	Hidrológico	688666150
União dos Palmares	AL	18/06/2010	Enxurradas	Hidrológico	212900000
Teresópolis	RJ	12/01/2011	Enxurradas	Hidrológico	150814075
Rio Branco	AC	05/02/2012	Inundações	Hidrológico	138927387
Canela	RS	01/11/2021	Estiagem e Seca	Climatológico	136000623
Congonhas	MG	02/01/2012	Inundações	Hidrológico	118719000
Boa Vista do Ramos	AM	03/05/2012	Inundações	Hidrológico	102000000
Angra dos Reis	RJ	06/01/2010	Movimento de Massa	Hidrológico	100709000

Figura 39 – 10 eventos de maior total de danos humanos

4 - Esses eventos coincidem?

Conforme as análises realizadas e apresentadas na Figura 33, Figura 36 e Figura 38, o evento com maio número de óbitos corresponde ao evento com maiores danos humanos, porém não coincide com o evento com maior prejuízo total.

Os 10 eventos com maior prejuízo total, que inclui prejuízos públicos e privados, não correspondem a nenhum dos 10 eventos com maior número de óbitos nem aos 10 eventos com maiores danos humanos, vide Figura 34, Figura 37 e Figura 39.

Porém, de acordo com a Figura 34 e Figura 39, alguns eventos de maior número de óbitos correspondem a eventos de maior total de danos humanos, a saber, Nova Friburgo (2011), Jaboatão dos Guararapes (2022), Teresópolis (2011) e Angra dos Reis (2010).

5 - Quais os municípios com maior ocorrência desse tipo de evento?

Considerando-se o evento de maiores danos humanos total, buscou-se, graficamente, os municípios com maior ocorrência. Conforme observado na Figura 40, o município de Nova Friburgo aparece em primeiro, seguido de Maceió, Jaboatão dos Guararapes, União dos Palmares e Teresópolis, Rio Branco, Canela, Congonhas, Boa Vista do Ramos e Angra dos Reis.

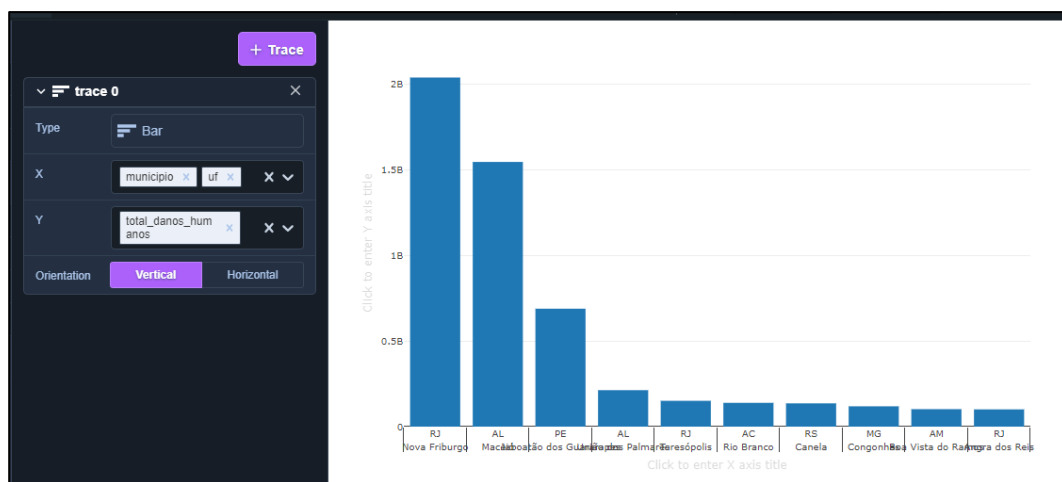


Figura 40 – Municípios com maior total de danos humanos

6 - Esses municípios pertencem ao mesmo estado brasileiro?

Ao analisar a ocorrência dos maiores danos humanos totais por estado, observa-se a presença de 7 estados, sendo estes: Rio de Janeiro, Alagoas, Pernambuco, Acre, Rio Grande do Sul, Minas Gerais e Amazonas, em ordem decrescente. Observa-se, ainda, de acordo com a Figura 41, que o estado do Rio de Janeiro concentra três municípios e o estado do Alagoas dois municípios.

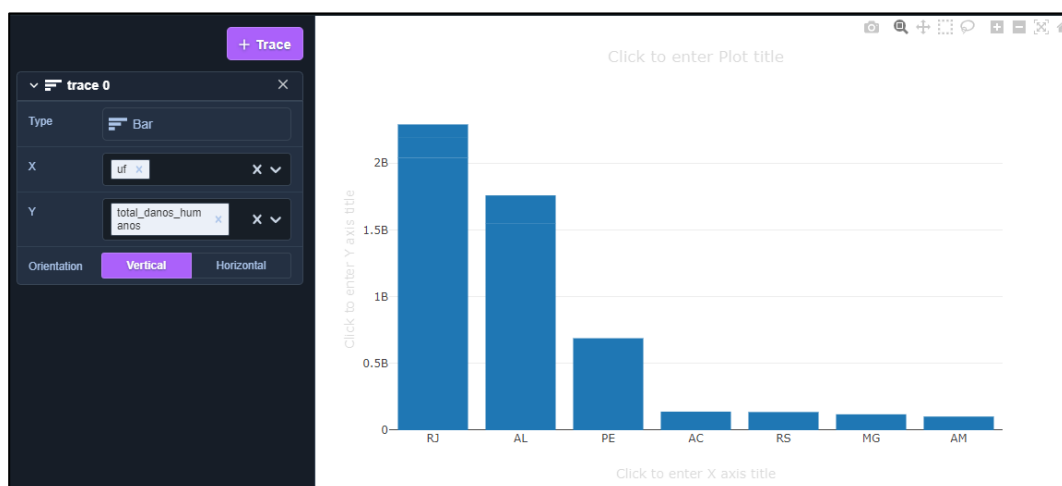
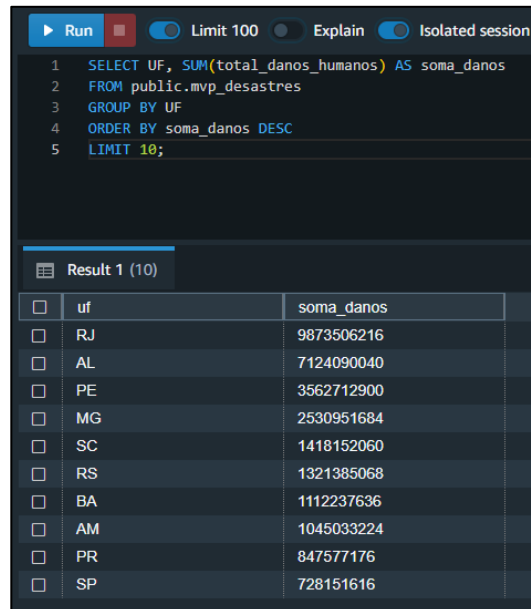


Figura 41 – Maiores totais de danos humanos por estado

7 - Caso contrário, qual estado brasileiro possui predominância desses eventos?

Corroborando o observado graficamente, foi feita a consulta, numérica dos 10 estados com maior ocorrência de danos humanos totais, constante no banco de dados. Conforme Figura 42 o estado com maior ocorrência de danos humanos é o estado do Rio de Janeiro, seguido pelo estado de Alagoas.



```
1 SELECT UF, SUM(total_danos_humanos) AS soma_danos
2 FROM public.mvp_desastres
3 GROUP BY UF
4 ORDER BY soma_danos DESC
5 LIMIT 10;
```

uf	soma_danos
RJ	9873506216
AL	7124090040
PE	3562712900
MG	2530951684
SC	1418152060
RS	1321385068
BA	1112237636
AM	1045033224
PR	847577176
SP	728151616

Figura 42 – Predominância de totais de danos humanos por estado

8 - É possível prever a qual fator natural esse evento pode ser associado?

Analisando-se, graficamente, a tipologia, e os grupos de desastres às quais pertencem, os eventos de maiores danos humanos totais, observa-se uma predominância nos eventos Hidrológicos, seguido por eventos Climatológicos, conforme apresentado na Figura 43. Entre os eventos hidrológicos, os de maior ocorrência são as enchurradas, seguidas por movimentos de massa, chuvas intensas e inundações. Enquanto os eventos climatológicos consistem em estiagem e seca.

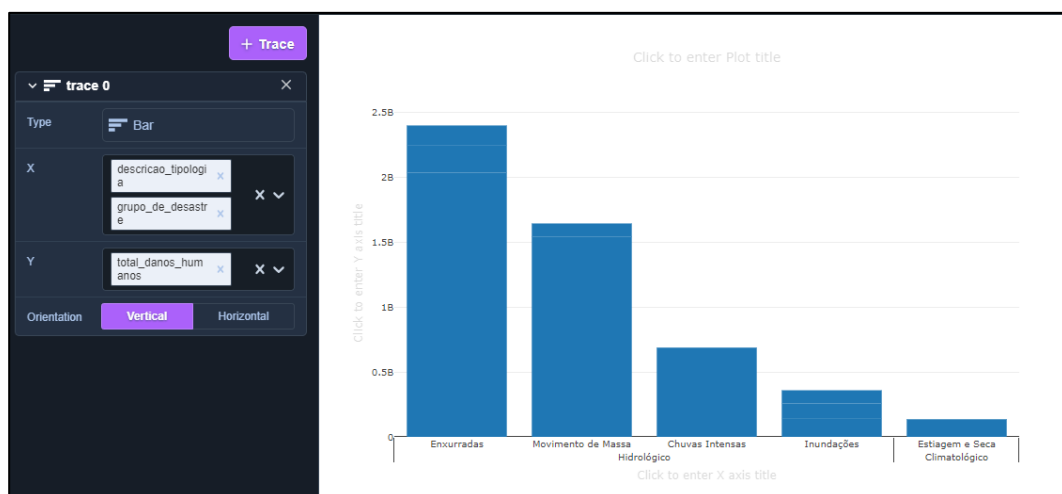


Figura 43 – Fatores naturais associados aos maiores totais de danos humanos

Conclusão

Por meio do presente trabalho foi possível trabalhar a construção de um pipeline de dados, envolvendo a busca, coleta, modelagem, carga e análise dos dados, utilizando tecnologia na nuvem da AWS.

O objetivo do trabalho foi avaliar a ocorrência de desastres naturais no Brasil, assim como o número de óbitos, prejuízo financeiro e total de danos humanos, associados a estes eventos, assim como fazer uma análise da ocorrência destes por municípios e estados brasileiros e correlacioná-los aos fatores naturais que os originaram.

Por meio das análises realizadas foi possível constatar que o evento com maior número de óbitos corresponde ao evento com maiores danos humanos, porém não coincide com o evento com maior prejuízo total. Os 10 eventos com maior prejuízo total, que inclui prejuízos públicos e privados, não correspondem a nenhum dos 10 eventos com maior número de óbitos nem aos 10 eventos com maiores danos humanos. Observou-se ainda, que a tipologia e os grupos de desastres correspondentes aos eventos de maiores danos humanos totais, são os eventos hidrológicos, sendo os de maior ocorrência as enxurradas, seguidas por movimentos de massa, chuvas intensas e inundações.

Autoavaliação

O objetivo proposto no início do presente trabalho foi alcançado, assim como foram respondidas as questões inicialmente propostas.

Porém, foram encontradas algumas dificuldades no início da execução do presente trabalho. Foram realizados testes iniciais nas três nuvens sugeridas para realização do trabalho, AWS, Microsoft Azure e Google Cloud Storage. Na Microsoft Azure foi encontrada uma dificuldade inicial pois a conta automaticamente foi vinculada ao diretório de uma empresa, sem sucesso de desconexão, até o presente momento. Na Google Cloud Storage, foi obtido um erro ao realizar a carga dos dados. Porém, ao corrigir os dados iniciais, optou-se por seguir o trabalho com a AWS, devido ao oferecimento de um tutorial mais detalhado e a uma interface mais amigável de trabalho.

Considerando-se que os grupos de desastre de maior ocorrência observados foram as enxurradas, seguidas por movimentos de massa, chuvas intensas e inundações, sugere-se, para trabalhos futuros, relacionar os dados utilizados no presente trabalho com dados pluviométricos e geotécnicos, de forma a enriquecer as análises realizadas e, eventualmente, estabelecer uma relação destes parâmetros com a ocorrência dos desastres, e fornecer subsídios para elaboração de um plano de prevenção para os mesmos.