



MVP – Disciplina: Sprint: Engenharia de Dados

Nome: Geam Piero Morales

Fonte: <https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results>

120 Anos de História dos Jogos Olímpicos: Atletas e Resultados

Esse é o Dataset dos jogos olímpicos da era moderna, indo de 1896 até 2016, no Rio de Janeiro

Observação: Os jogos de verão e inverno ocorriam no mesmo ano até 1992, onde foram divididos em uma diferença de 2 anos entre eles

Objetivo desse MVP é a criação de um pipeline de engenharia de dados onde podemos analisar graficamente a relação entre a evolução dos jogos

Esse Dataset traz informações sobre um compilado desde 1900 até 2016 com nome, países, medalhas, identificação, idade e outros. Essas informações servem para contar a história dos participantes e vencedores.

Além dessas questões acima, podemos enxergar possíveis evoluções para o futuro.

A partir disso, será importado esse Dataset para o AWS, onde também será realizado o ETL também, primeiramente utilizando o S3, indo na sequência para o Glue e finalizando no Redshift

Qualidade dos Dados:

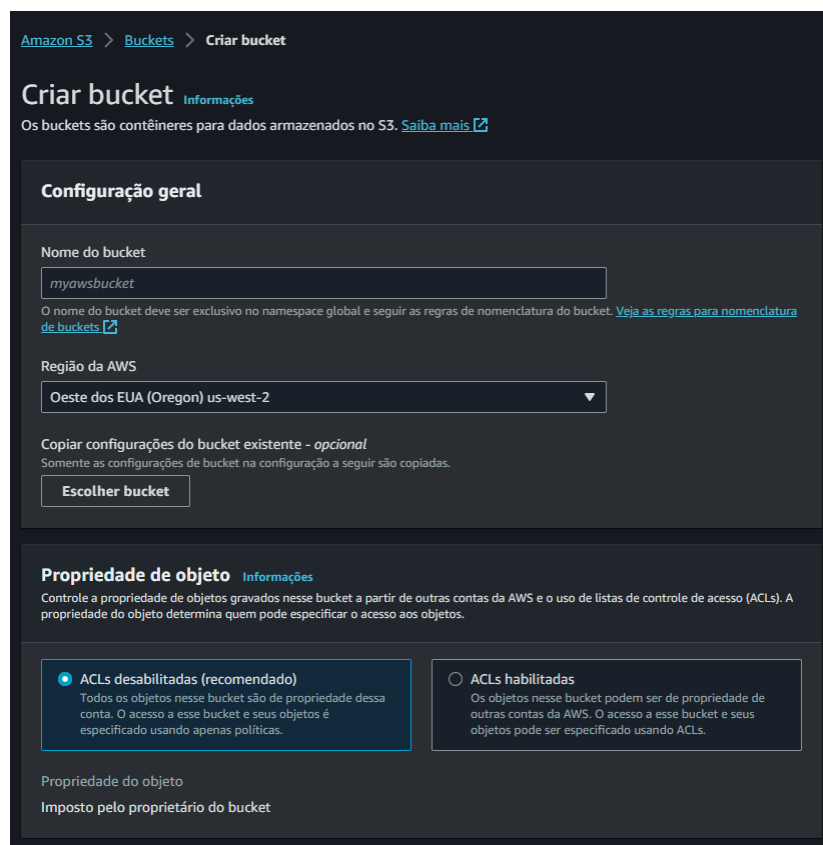
Feita verificação, os dados se encontram sem problemas para visualização, não havendo dados nulos, duplicados ou ausente.

1º Etapa: Criação do Bucket (AWS S3)

Definição:

Amazon S3 (Simple Storage Service) é um “bucket” é um contêiner de armazenamento que permite armazenar e organizar dados nuvem da AWS, sendo necessário a criação de um *bucket* para armazenamento.

1.1 – Clicar em Bucket e ir em “Criar Bucket”



Amazon S3 > Buckets > Criar bucket

Criar bucket Informações

Os buckets são contêineres para dados armazenados no S3. [Saiba mais](#)

Configuração geral

Nome do bucket

myawsbucket

O nome do bucket deve ser exclusivo no namespace global e seguir as regras de nomenclatura do bucket. [Veja as regras para nomenclatura de buckets](#)

Região da AWS

Oeste dos EUA (Oregon) us-west-2

Copiar configurações do bucket existente - *opcional*

Somente as configurações de bucket na configuração a seguir são copiadas.

[Escolher bucket](#)

Propriedade de objeto Informações

Controle a propriedade de objetos gravados nesse bucket a partir de outras contas da AWS e o uso de listas de controle de acesso (ACLs). A propriedade do objeto determina quem pode especificar o acesso aos objetos.

☒ **ACLs desabilitadas (recomendado)**

Todos os objetos nesse bucket são de propriedade dessa conta. O acesso a esse bucket e seus objetos é especificado usando apenas políticas.

☐ **ACLs habilitadas**

Os objetos nesse bucket podem ser de propriedade de outras contas da AWS. O acesso a esse bucket e seus objetos pode ser especificado usando ACLs.

Propriedade do objeto

Imposto pelo proprietário do bucket

Ao colocar um nome de seu *bucket*, sendo necessário colocar uma região AWS. Além disso, devemos colocar as propriedades de objeto

O acesso público é concedido a buckets e objetos por meio de listas de controle de acesso (ACLs), políticas de bucket, políticas de ponto de acesso ou todas elas. Para garantir que o acesso público a este bucket e todos os seus objetos seja bloqueado, ative a opção de Bloquear todo o acesso público. Essas configurações serão aplicadas apenas a este bucket e aos respectivos pontos de acesso. A AWS recomenda ativar a opção Bloquear todo o acesso público. Porém, antes de aplicar qualquer uma dessas configurações, verifique se as aplicações funcionarão corretamente sem acesso público. Caso precise de algum nível de acesso público a este bucket ou aos objetos que ele contém, é possível personalizar as configurações individuais abaixo para que atendam aos seus casos de uso de armazenamento específicos. [Saiba mais](#)

Ativar essa configuração é o mesmo que ativar todas as quatro configurações abaixo. Cada uma das configurações a seguir são independentes uma da outra.

- O versionamento é um meio de manter múltiplas variantes de um objeto no mesmo bucket. Você pode usar o versionamento para preservar, recuperar e restaurar todas as versões de cada objeto armazenado no bucket do Amazon S3. Com o versionamento, você pode recuperar facilmente ações não intencionais do usuário e falhas da aplicação. [Saiba mais](#)

Você pode usar tags de bucket para rastrear custos de armazenamento e organizar buckets. [Saiba mais](#).

Nenhuma tag associada a este bucket.

Adicionar tag

A criptografia no lado do servidor é aplicada automaticamente a novos objetos armazenados nesse bucket.

- **Criptografia do lado do servidor com chaves gerenciadas do Amazon S3 (SSE-S3)**

- Chave do bucket**
- O uso de uma chave de bucket do S3 para SSE-KMS reduz os custos de criptografia ao diminuir as chamadas para o AWS KMS. As chaves de bucket do S3 não são compatíveis com o DSSE-KMS. [Saiba mais](#)

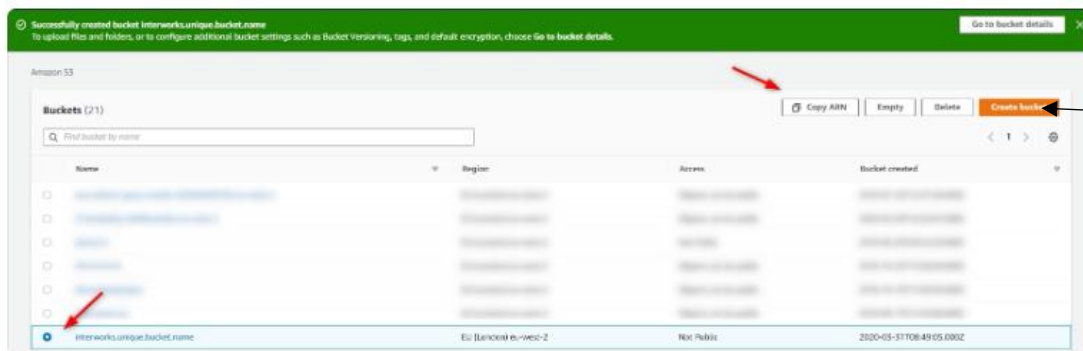
- ☐ Desativar
- ☒ Ativar

► **Configurações avançadas**

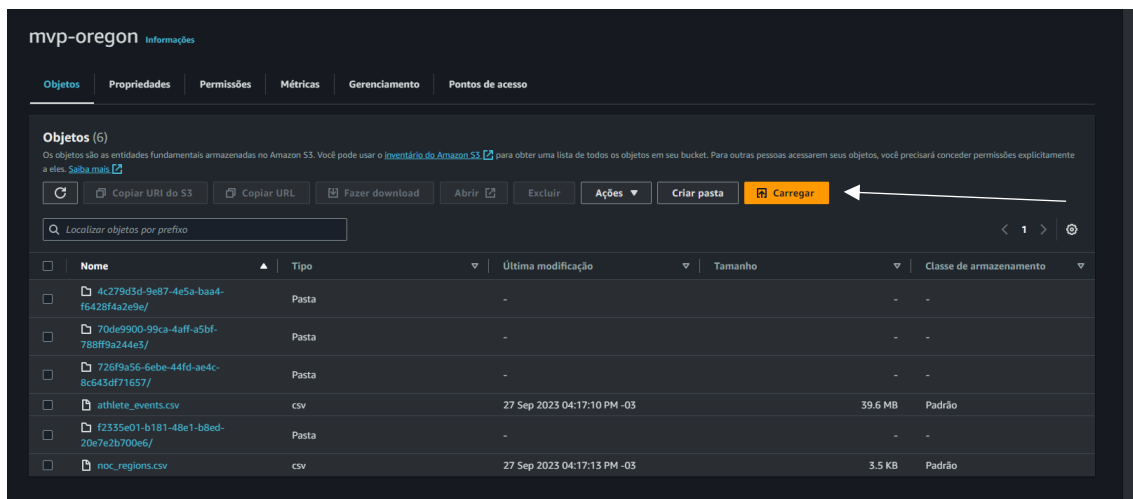
Depois de criar o bucket, você pode fazer upload de arquivos e pastas para o bucket e definir configurações adicionais do bucket.

Cancelar

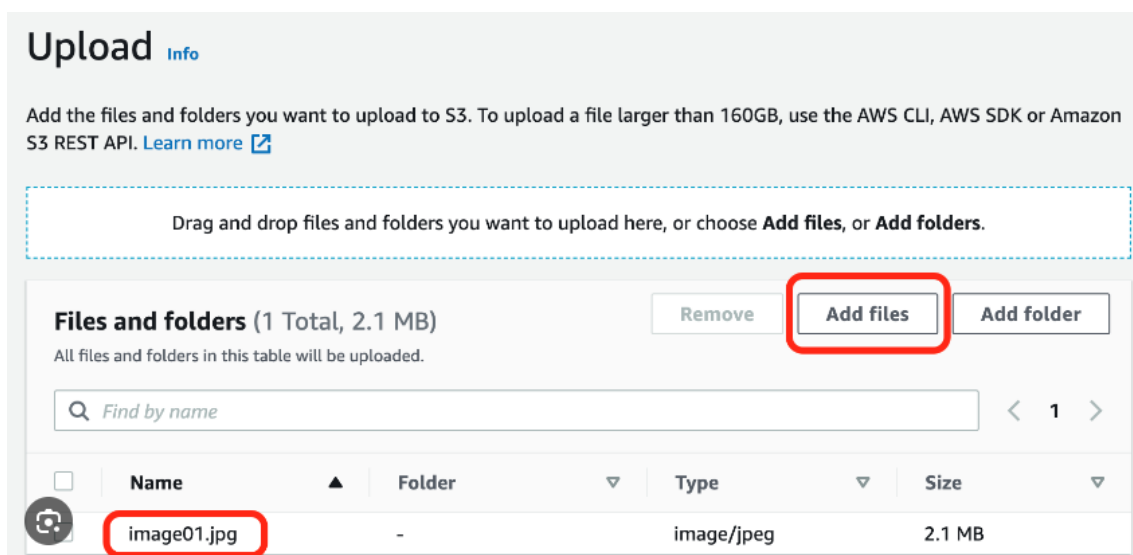
Criar bucket



Após isso, vamos em “Carregar” para adicionarmos os arquivos



Pode-se adicionar tanto arquivo e uma pasta inteira



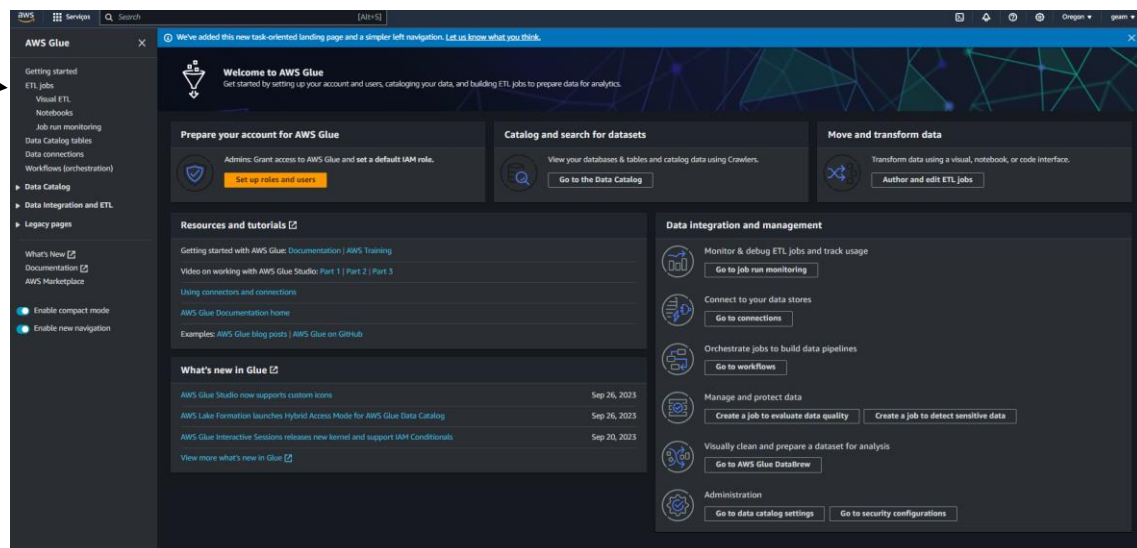
Após isso, podemos dar o upload dos arquivos CSV utilizados para as análises.

2º Etapa: Criação do Glue (AWS Glue)

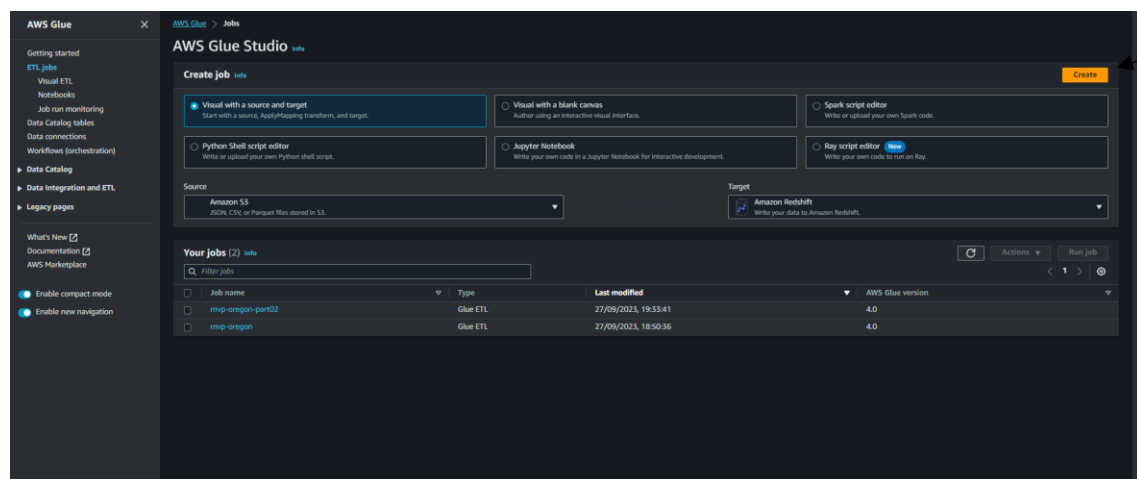
Descrição:

O AWS Glue é um serviço oferecido pela Amazon para realização de recursos de ETL (Extração, Transformação e Carga) em nuvem. Possuindo capacidade para automatizar tarefas de integração e transformação, assim, aumentando a eficiência de seus usuários.

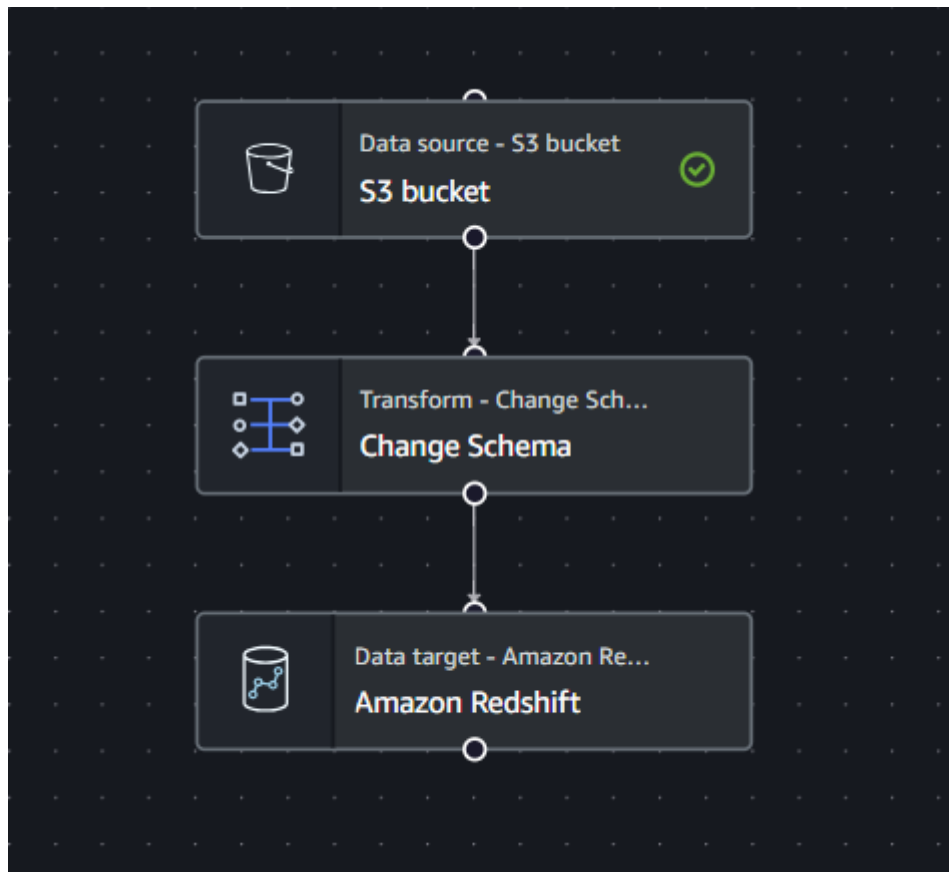
Como premissa de simplificar os processos de ETL, acaba acelerando o processo de geração de insights e diminuindo o tempo do usuário em fazer o gerenciamento de dados.



Para criar um ETL, devemos clicar em “ETL Jobs”



Para criar um Glue do zero, devemos clicar em “Criar”



Devemos configurar dentro do Glue, onde está o S3, dando a informação da URL e o caminho para que o Glue consiga localizar onde está armazenado. Além disso, deve-se já dizer que tipo de dado é, no caso CSV e identificar o separador

Data source properties - S3
Output schema
Data preview

Name
S3 bucket

S3 source type
Info
☒ S3 location
Choose a file or folder in an S3 bucket.
☐ Data Catalog table

S3 URL

☒ Recursive
Read files in all subdirectories.

Data format
CSV

Delimiter
Comma (,)

Escape character - optional
Enter a character to use for escaping

The character which immediately follows is used as-is, except for a small set of well-known escapes (\n, \r, \t, and \0)

Quote character
Double quote (")

☒ First line of source file contains column headers
☐ Records in source files can span multiple lines

▶ Additional options

Para realizarmos um ETL, devemos aplicar um “*Change Schema*”, permite realizar o tratamento das informações e colocar de onde os dados estão em “*Node Parents*”

Transform
Output schema
Data preview

Name
Change Schema

Node parents
Choose which nodes will provide inputs for this one.
Choose one or more parent node

S3 bucket
S3 - DataSource

Change Schema (Apply mapping)

Source key	Target key	Data type	Drop
ID	ID	string	<input type="checkbox"/>
Name	Name	string	<input type="checkbox"/>
Sex	Sex	string	<input type="checkbox"/>
Age	age	int	<input type="checkbox"/>
Height	Height	string	<input type="checkbox"/>
Weight	Weight	string	<input type="checkbox"/>
Team	Team	string	<input type="checkbox"/>
NOC	NOC	string	<input type="checkbox"/>
Games	Games	string	<input type="checkbox"/>
Year	Year	string	<input type="checkbox"/>
Season	Season	string	<input type="checkbox"/>
City	City	string	<input type="checkbox"/>
Sport	Sport	string	<input type="checkbox"/>
Event	Event	string	<input type="checkbox"/>
Medal	Medal	string	<input type="checkbox"/>

3º Etapa: Criação do AWS Redshift

Descrição:

O AWS Redshift é um serviço de armazenamento e análises de Data Warehouse com alta escala, sendo utilizado para consumo e criação de comandos SQL para uma volumetria de dados de alta escala e geração de análise. Logo, para criarmos um banco, vamos seguir o passo-passo abaixo:

[Amazon Redshift Serverless](#) > Primeiros passos com o Amazon Redshift Serverless

Primeiros passos com o Amazon Redshift Serverless [Informações](#)

Para começar a usar o Amazon Redshift Serverless, configure seu data warehouse sem servidor e crie um banco de dados. Você receberá USD 298,30 de crédito para o uso do Redshift Serverless nesta conta.

Configuração

☒ Usar configurações padrão
As configurações padrão foram definidas para ajudar você a começar. É possível alterá-las a qualquer momento mais tarde.

☐ Personalizar configurações
Personalize suas configurações de acordo com suas necessidades específicas.

Namespace [Informações](#)

Namespace is a collection of database objects and users. Data properties include database name and password, permissions, and encryption and security.

⚠ Seus dados são criptografados por padrão com uma chave de propriedade da AWS. Para escolher uma chave diferente, escolha **Personalizar configurações**.

Target namespace
default-namespace

Nome e senha do banco de dados

Nome do banco de dados dev	Credenciais do usuário administrador Credenciais do IAM fornecidas
-------------------------------	---

Permissões

Devemos também criar um usuário administrador para que consiga realizar as políticas de uso do banco de dados

Create the default IAM role

Create an IAM role as the default for this cluster that has the [AmazonRedshiftAllCommandsFullAccess](#) policy attached. This policy includes permissions to run SQL commands to COPY, UNLOAD, and query data with Amazon Redshift. The policy also grants permissions to run SELECT statements for related services, such as Amazon S3, Amazon CloudWatch logs, Amazon SageMaker, and AWS Glue.

Specify an S3 bucket for the IAM role to access

To create a new bucket, [\[object Object\]](#)

☐

No additional S3 bucket

Create the IAM role without specifying S3 buckets.

☒

Any S3 bucket

Allow users that have access to your Redshift cluster to also access any S3 bucket and its contents in your AWS account.

☐

Specific S3 buckets

Specify one or more S3 buckets that the IAM role being created has permission to access.

Cancel

Create IAM role as default

Cluster permissions

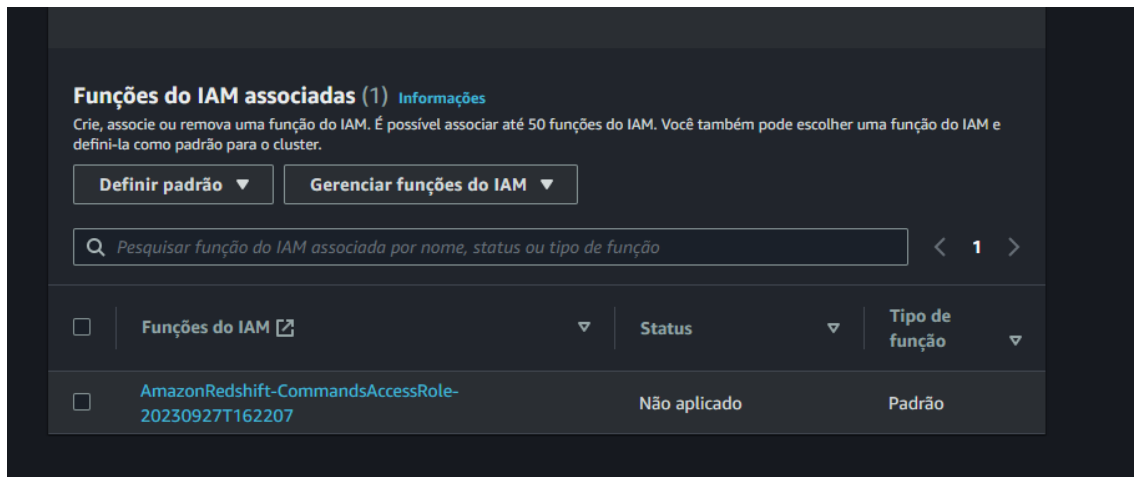
Create an IAM role as the default for this cluster that has the [AmazonRedshiftAllCommandsFullAccess](#) policy attached. This policy includes permissions to run SQL commands to COPY, UNLOAD, and query data with Amazon Redshift. The policy also grants permissions to run SELECT statements for related services, such as Amazon S3, Amazon CloudWatch logs, Amazon SageMaker, and AWS Glue.

Manage IAM roles

Create, associate, or remove an IAM role. You can associate up to 50 IAM roles. You can also choose an IAM role and set it as the default for this cluster.

The IAM role [AmazonRedshift-CommandsAccessRole-20211122T151232](#) was successfully created and set as the default for this cluster.

Associated IAM roles (2) Info			
<div> <div>Set default</div> <div>Manage IAM roles</div> </div>		<div> <div>Search for associated IAM role by name, status, or role type</div> <div>< 1 ></div> </div>	
<input type="checkbox"/>	IAM roles ↗	Status	Role type ▲
<input type="checkbox"/>	AmazonRedshift-CommandsAccessRole-20211119T025247	Not applied	--



Após a criação da IAM, devemos também configurar a rede, mantendo também o VPC e Subnet

Target namespace
default-namespace

Nome e senha do banco de dados

Nome do banco de dados
dev

Credenciais do usuário administrador
Credenciais do IAM fornecidas

Permissões

Função do IAM padrão
arn:aws:iam::980793441388:role/service-role/AmazonRedshift-CommandsAccessRole-20230927T162207

Criptografia e segurança

Criptografia do AWS KMS
Chave do KMS de propriedade da AWS

Registro em log de auditoria
Desativado

Grupo de trabalho [Informações](#)

Workgroup is a collection of compute resources from which an endpoint is created. Compute properties include network and security settings.

Workgroup name
default-workgroup

Capacidade básica em unidades de processamento (RPU) do Redshift [Informações](#)

The capacity is measured in Redshift processing units (RPU).

Capacidade básica de RPU
128

Rede e segurança

Virtual Private Cloud (VPC)
vpc-045a82d1b9bf10f93

Sub-rede
subnet-039b296f5552cf026,
subnet-065e6cdad3be3293c,
subnet-090498dad90bf82ac,
subnet-031797e061b2d311c

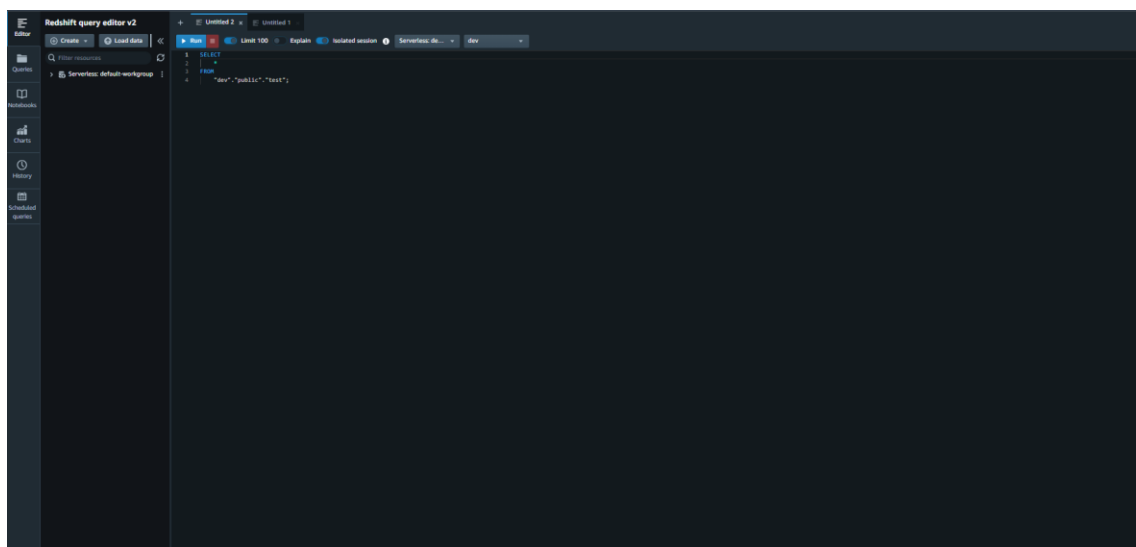
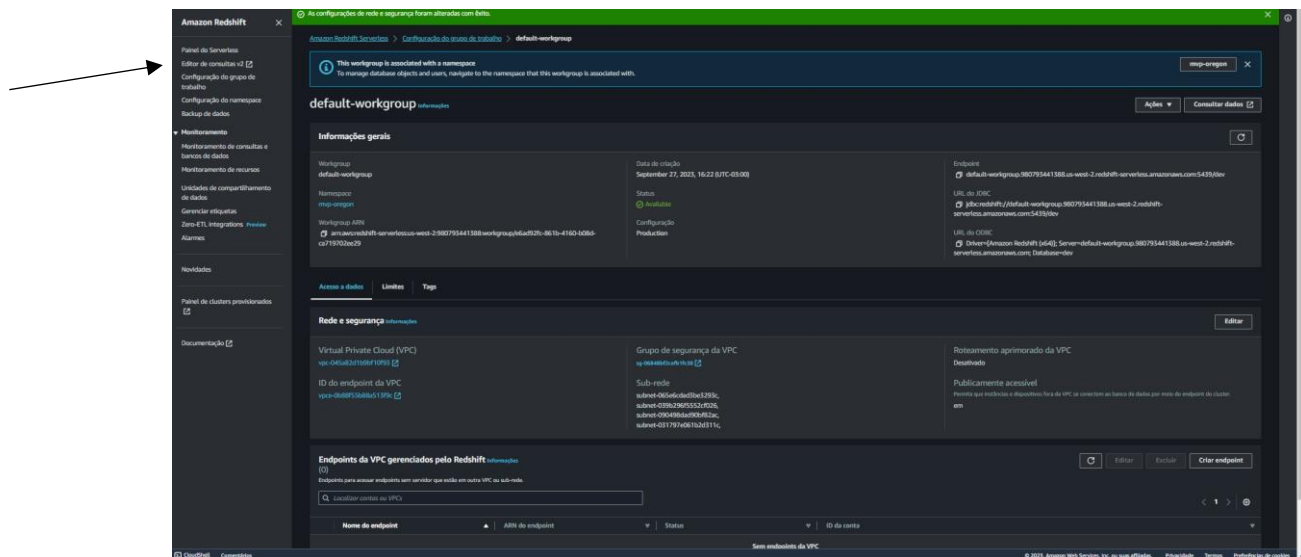
Grupo de segurança da VPC
sg-06848bf2cafb1fc38

Roteamento aprimorado da VPC
Desativado

[Cancelar](#) [Salvar configuração](#)

Comentários

Depois, irá para a página inicial, onde será dados as tabelas, para isso, iremos em “Editar Consultas V2”



Voltando para o AWS Glue para configurarmos a conexão entre o Glue e o Redshift

Name
Enter a unique name for your connection.

Connection type

☐ **Require SSL connection**
The connection will fail if it's unable to connect over SSL.

Description - optional

Descriptions can be up to 2048 characters long.

Connection access

Database instances
Provisioned Amazon Relational Database Service instances.

Database name

Credential type
☒ Username and password
☐ AWS Secrets Manager

Username

Além disso, deve-se utilizar a criação também um *endpoint*, para isso, devemos acessar o “VPC” e selecionar o “Endpoint”

Painel da VPC
Visualização global do VPC ☒ Novo
Filtrar por VPC:

Novo privado virtual
 Suas VPCs [Novo](#)
 Sub-redes
 Tabelas de rotas
 Gateways da Internet
 Gateways da Internet somente de saída
 Gateways da operadora
 Conjuntos de opções de DHCP
 IPs elásticos
 Listas de prefixos gerenciados
 Endpoints
 Serviços de endpoint
 Gateways NAT
 Conexões de empastamento

Segurança
 ACLs da rede
 Grupos de segurança
Firewall de DNS
 Grupos de regras
 Listas de domínios
Network Firewall
 Firewall
 Políticas de firewall

Recursos por região
 Você está usando os seguintes recursos da Amazon VPC

VPCs Ver todas as regiões	Gateways NAT Ver todas as regiões
Sub-redes Ver todas as regiões	Conexões de empastamento de VPC Ver todas as regiões
Tabelas de rotas Ver todas as regiões	Network ACLs Ver todas as regiões
Gateways da Internet Ver todas as regiões	Grupos de segurança Ver todas as regiões
Gateways da Internet somente de saída Ver todas as regiões	Gateways do cliente Ver todas as regiões
Conjuntos de opções de DHCP Ver todas as regiões	Gateways privados virtuais Ver todas as regiões
IPs elásticos Ver todas as regiões	Conexões VPN site a site Ver todas as regiões
Endpoints Ver todas as regiões	Instâncias em execução Ver todas as regiões
Serviços de endpoint Ver todas as regiões	

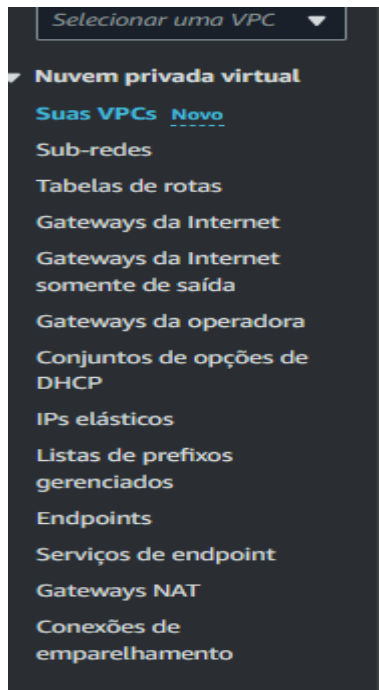
Integridade de serviço
 Visualizar todos os detalhes da integridade do serviço [↗](#)

Configurações
 Zonas
 Experimentos de console

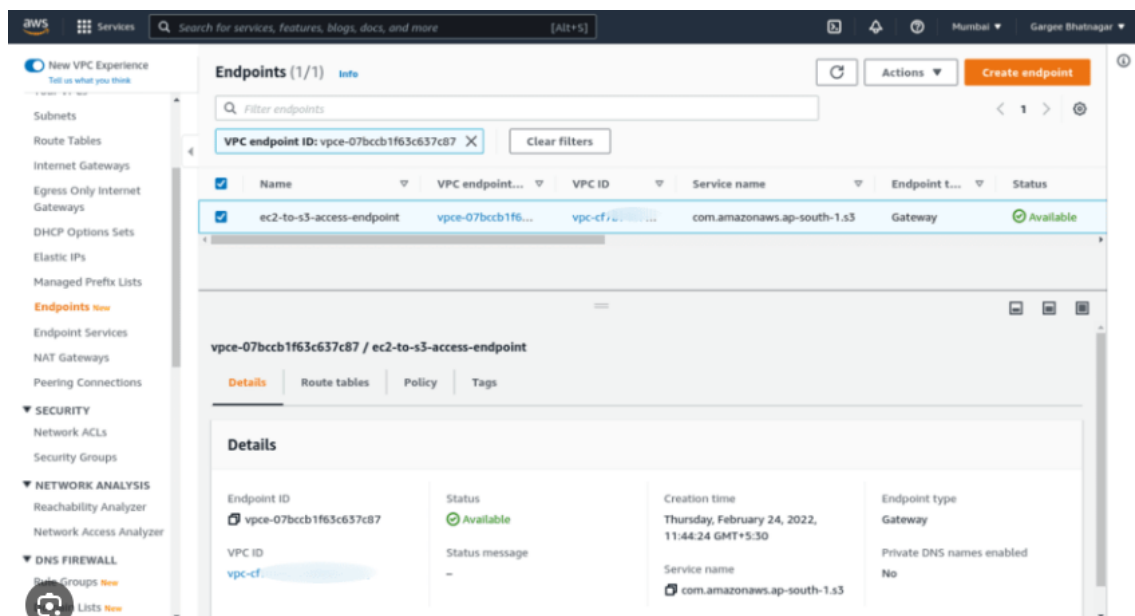
Informações adicionais
 Documentação da VPC
 Todos os recursos da VPC
 Fóruns
 Relatar um problema

AWS Network Manager
 O AWS Network Manager fornece ferramentas e recursos para ajudar você a gerenciar e monitorar sua rede na AWS. O Network Manager facilita a execução de gerenciamento de conectividade, monitoramento e solução de problemas de rede, gerenciamento de IP e segurança e governança de rede.
 Comece a usar o Network Manager

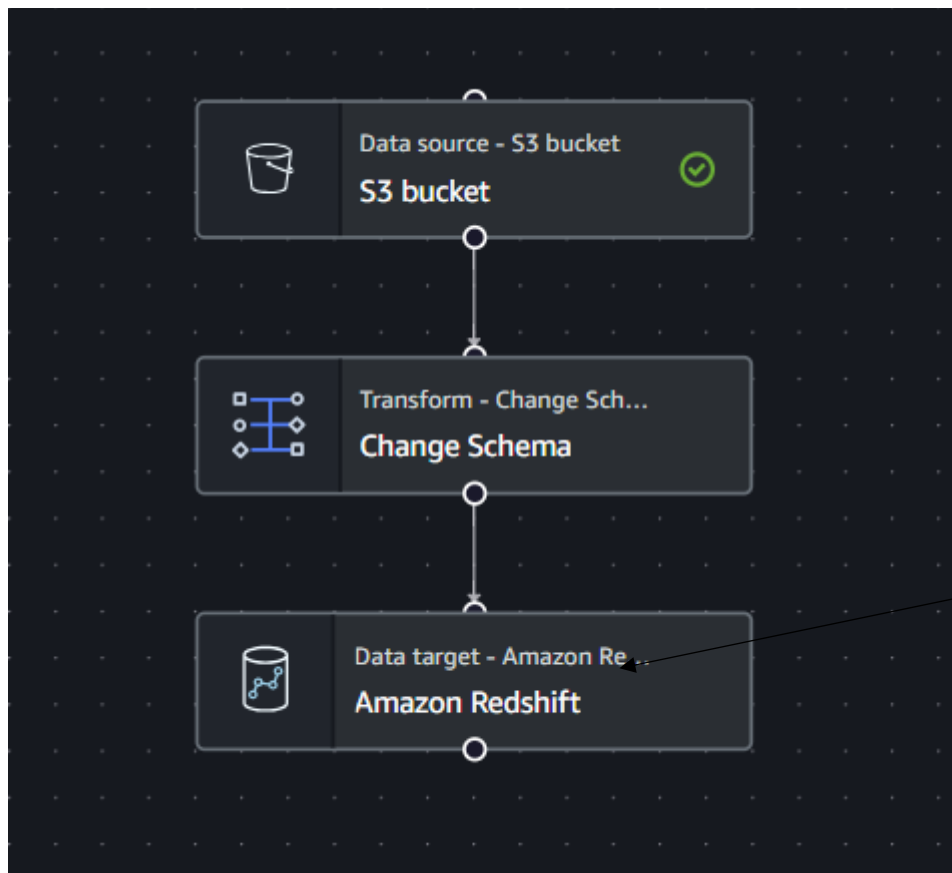
Conexões VPN site a site
 A Amazon VPC permite que você use seus próprios recursos isolados dentro da Nuvem AWS e, em seguida, conecte esses recursos diretamente ao seu próprio datacenter usando padrão do setor de conexões criptografadas IPsec VPN.



Selecionamos “Create Endpoint”



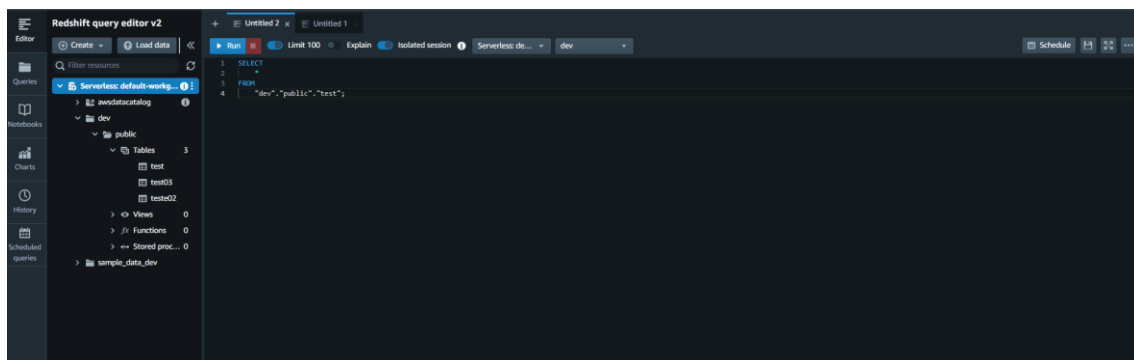
Após as configurações, retornamos ao AWS Glue e selecionamos “Amazon Reshift”



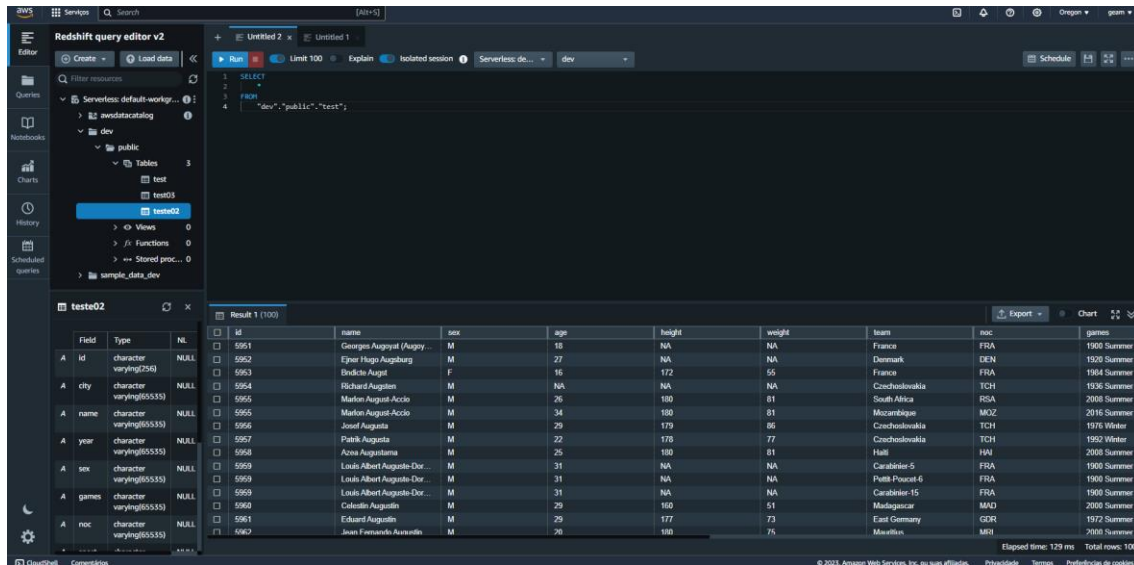
Após, clicar na configuração da Amazon Redshift

The screenshot shows the 'Data target properties - Amazon Redshift' configuration window. The 'Name' field is 'Amazon Redshift'. The 'Node parents' dropdown is set to 'Choose one or more parent node'. The 'Redshift access type' is 'Direct data connection - recommended'. The 'Redshift connection' is 'mvp-oregon'. The 'Connection' is 'dev'. The 'Schema' is 'public'. The 'Table' is 'teste02'. The 'Handling of data and target table' is 'TRUNCATE target table'. A warning message states: 'Truncate will erase all data in the target table with every job run.'

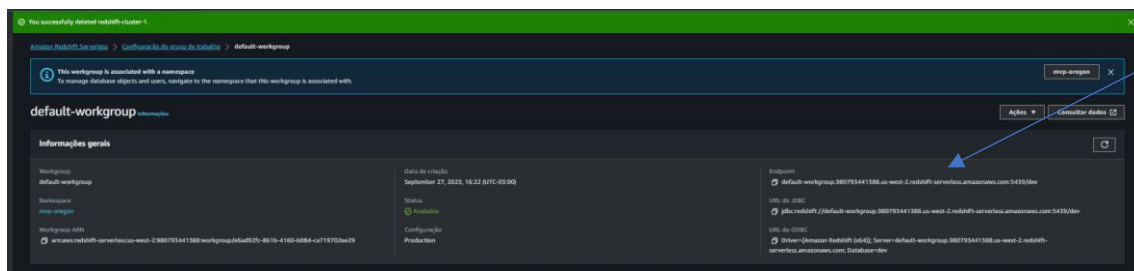
Após dentro do Glue – Amazon Redshift, devemos colocar o banco de dados que foi criado no Amazon Redshift. Além disso, devemos colocar a tabela “Public” no “Schema”



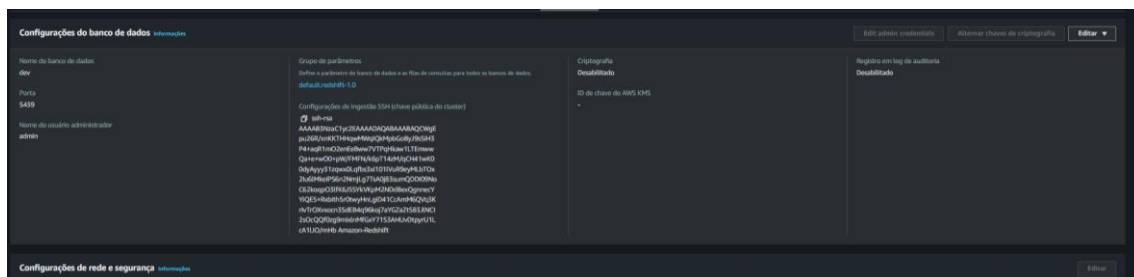
Após isso, devemos escrever o nome que deseja para gerar uma tabela que vai para o Amazon Redshift e selecionamos o “Truncate” para que consiga levar os dados para o banco de dados e selecionamos em “Save” e rodamos o “Run”



Após todas as configurações, iremos para o Microsoft Power BI, onde faremos a interpretação de dados, para isso devemos configurar o Amazon Redshift e buscamos por “EndPoint” para que possamos realizar a conexão no Microsoft Power BI



Para habilitarmos acesso do Power BI aos dados do Amazon Redshift, devemos ir em configurações de banco de dados e clicar em “Editar” e “Habilitar” em ambos, que irá permitir que o Power BI acesse os dados



No Power BI, irá solicitar as credenciais onde devemos colocar inicialmente o IP:

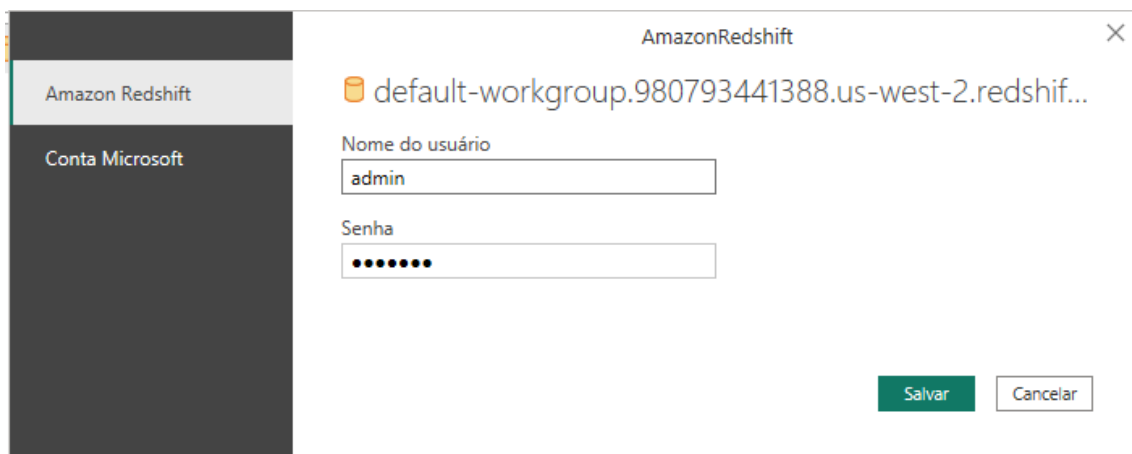
IP: default-workgroup.980793441388.us-west-2.redshift-serverless.amazonaws.com:5439

Servidor: dev

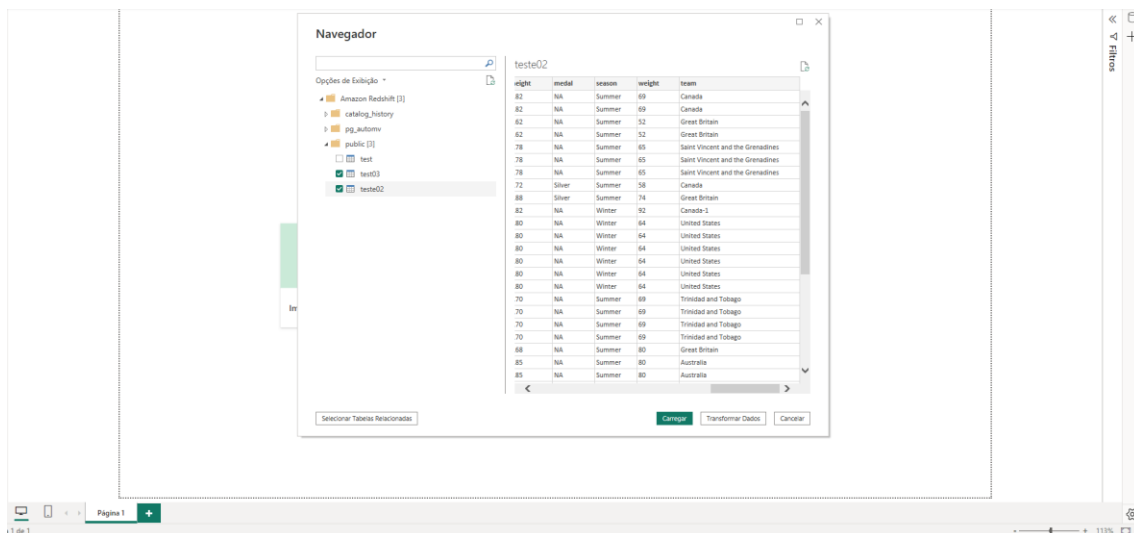
Colocar as credenciais do banco de dados da Amazon Redshift de login de administrador



A screenshot of the 'Amazon Redshift' connection dialog box. It has a title bar with a close button. Inside, there are two text input fields: 'Server' with the value 'default-workgroup.238559487473.us-east-1.redshift-serverless.amazonaws.com:5439' and 'Database' with the value 'dev'. Below these fields is a link that says 'Opções avançadas'. At the bottom right are two buttons: 'OK' and 'Cancelar'.



A screenshot of the Amazon Redshift login interface. On the left is a dark sidebar with 'Amazon Redshift' and 'Conta Microsoft' options. The main area has a title 'AmazonRedshift' and a close button. Below the title is the server address 'default-workgroup.980793441388.us-west-2.redshif...'. There are two input fields: 'Nome do usuário' with the value 'admin' and 'Senha' with masked characters. At the bottom right are two buttons: 'Salvar' and 'Cancelar'.



A screenshot of the Power BI interface. The 'Navegador' (Navigator) pane on the left shows a tree view with 'Amazon Redshift' expanded, showing 'teste02' selected. The main area displays a table with columns: 'id', 'medal', 'season', 'weight', and 'team'. The table contains 20 rows of data. At the bottom of the table are buttons: 'Selecionar Tabelas Relacionadas', 'Carregar', 'Transformar Dados', and 'Cancelar'.

id	medal	season	weight	team
82	NA	Summer	69	Canada
82	NA	Summer	69	Canada
62	NA	Summer	52	Great Britain
62	NA	Summer	52	Great Britain
78	NA	Summer	65	Saint Vincent and the Grenadines
78	NA	Summer	65	Saint Vincent and the Grenadines
78	NA	Summer	65	Saint Vincent and the Grenadines
72	Silver	Summer	58	Canada
88	Silver	Summer	74	Great Britain
82	NA	Winter	92	Canada-1
80	NA	Winter	64	United States
80	NA	Winter	64	United States
80	NA	Winter	64	United States
80	NA	Winter	64	United States
80	NA	Winter	64	United States
80	NA	Winter	64	United States
70	NA	Summer	69	Trinidad and Tobago
70	NA	Summer	69	Trinidad and Tobago
70	NA	Summer	69	Trinidad and Tobago
68	NA	Summer	80	Great Britain
85	NA	Summer	80	Australia
85	NA	Summer	80	Australia

Os dados para visualização e análise se encontra no link abaixo:

<https://app.powerbi.com/reportEmbed?reportId=85c6a58e-f49e-4748-a294-83869356a41b&autoAuth=true&ctid=13a9c019-a490-441f-b27d-6171403ea71f>

Nesse link, podemos encontrar evolução da entrada de atletas do sexo feminino após os jogos de 1992. Também podemos visualizar que os Estados Unidos é o país que mais envia atletas, além de ser a pessoa que mais possui medalhas.



