

Aluno: Miguel Costa Góes

Curso de pós graduação PUC RIO

Sprint 3 Engenharia de Dados

## Relatório do trabalho MVP da terceira Sprint

### Objetivo do trabalho:

O presente trabalho de pós-graduação em Ciência de Dados tem como principal meta a extração de insights a partir de conjuntos de dados obtidos por meio de uma abordagem de computação em nuvem. Neste contexto, descobrir qual são as empresas mais valorizadas da bolsa de valores, para isso a plataforma de nuvem escolhida para a execução deste projeto foi a Amazon Web Services (AWS).

### Esquema do trabalho

O trabalho foi organizado utilizando o esquema proposto em aula, com podemos ver no diagrama abaixo:

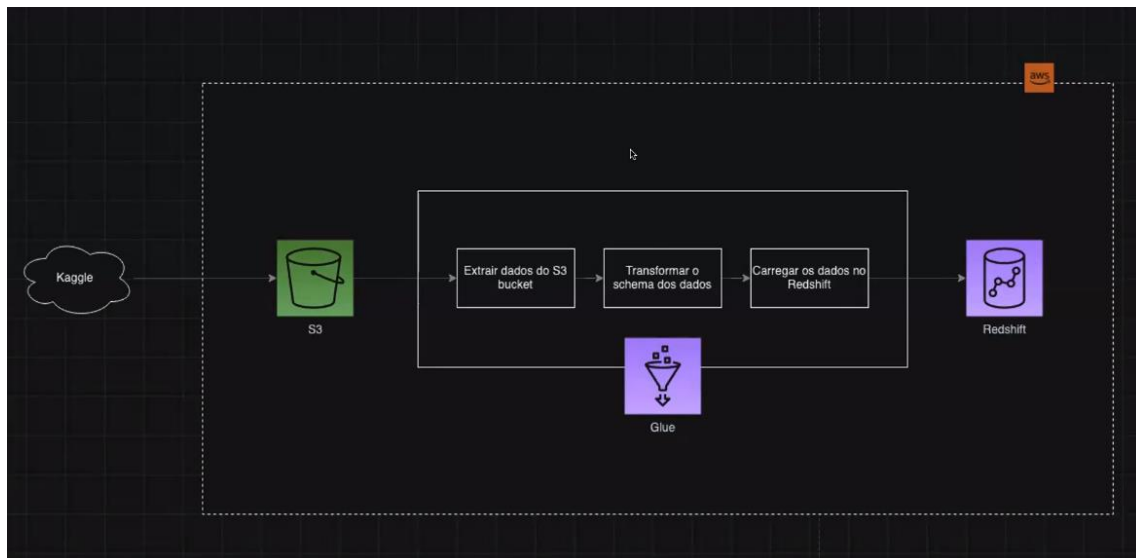


Figura 1 Esquema do trabalho

### Escolha coleta e tratamento dos dados

Para a realização do trabalho foi utilizado o dataset, Ações Brasileiras entre 2018 e 2021, esse data set pode ser encontrado no site do kaggle ou utilizando o link <https://www.kaggle.com/datasets/lucasprado/acoesbrasileiras2018a2021>.



*Figura 2 site kaggle onde foi retirado os dados*

Esse dataset possui um tamanho 44,03 MB e os dados com as seguintes colunas:

- 1- Preço da ação na abertura do pregão.
- 2- Maior preço que a ação alcançou;
- 3- Menor preço que a ação alcançou;
- 4- Preço da ação na fechamento do pregão;
- 5- volume transacionado;
- 6- Preço ajustado;
- 7- Data de referência do dado (Data).
- 8- Código da empresa na bolsa de valores;
- 9- Variação do preço ajustado da ação entre o dia em análise e o dia anterior;
- 10- Variação do preço ajustado de fechamento da ação entre o dia em análise e o dia anterior;
- 11- Nome da empresa em análise;
- 12- Setor de atuação da empresa;
- 13- Subsetor de atuação da empresa;
- 14- Tipo da ação (Tipo) (1\*BDR, 2\*ORDINARIA, 3\*REFERENCIAL)

15 - classificação pelo o faturamento da empresa;

Apesar dos datasets que serem tratados no site da kaggle foi realizado uma análise dos dados utilizando Python para garantir que não existia dados faltantes.

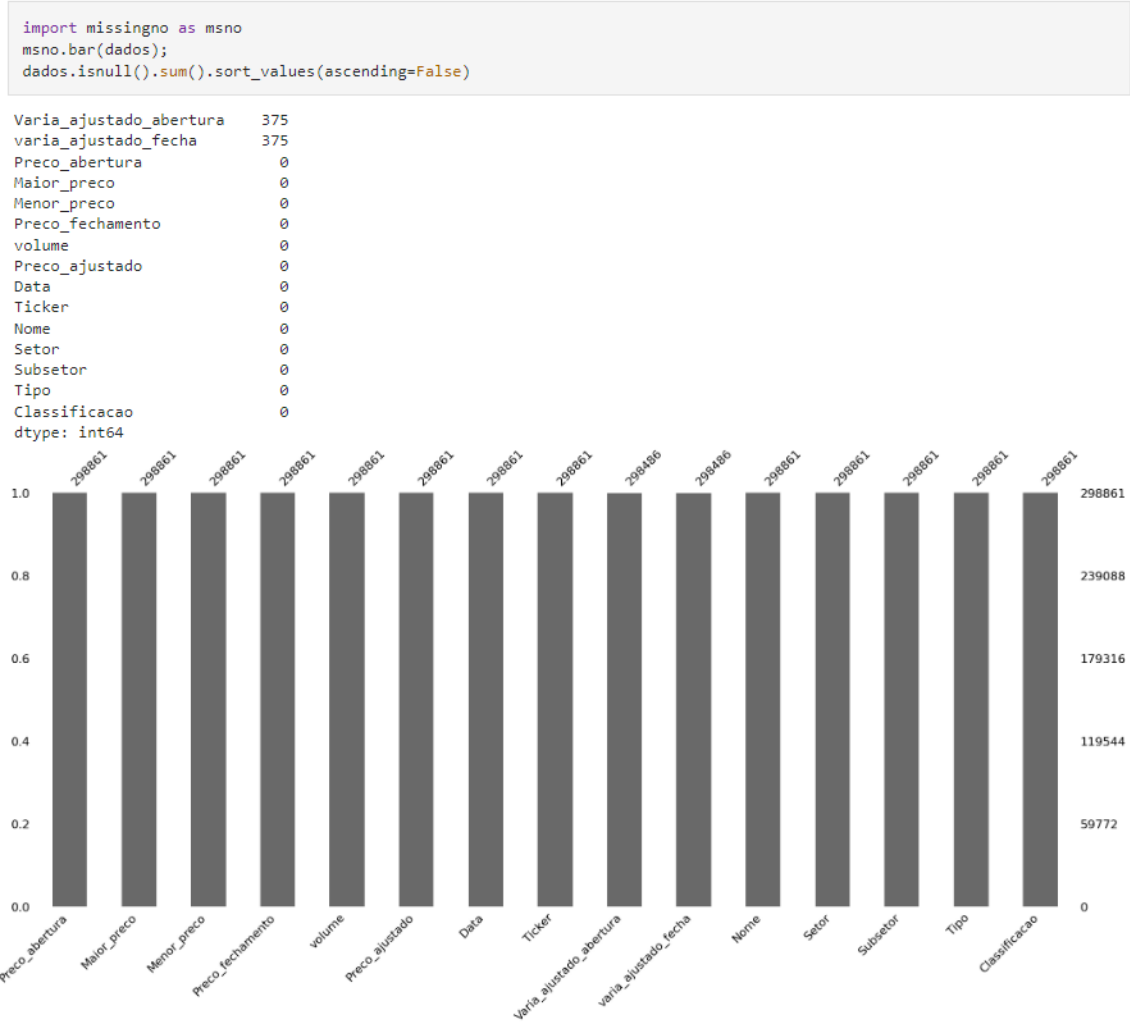


Figura 3 Verificação de dados faltantes

Como foram encontrados dados faltantes para algumas colunas, como é mostrado na imagem acima, foi necessário o tratamento prévio desses dados antes de carrega-los. O tratamento escolhido foi a exclusão das linhas desse dataset, pois os dados faltantes fazem referência a operações matemáticas com base em uma data fora do período de análise. Após a exclusão dessas linhas foi realizado outra verificação para garantir que todos os dados estão presentes.

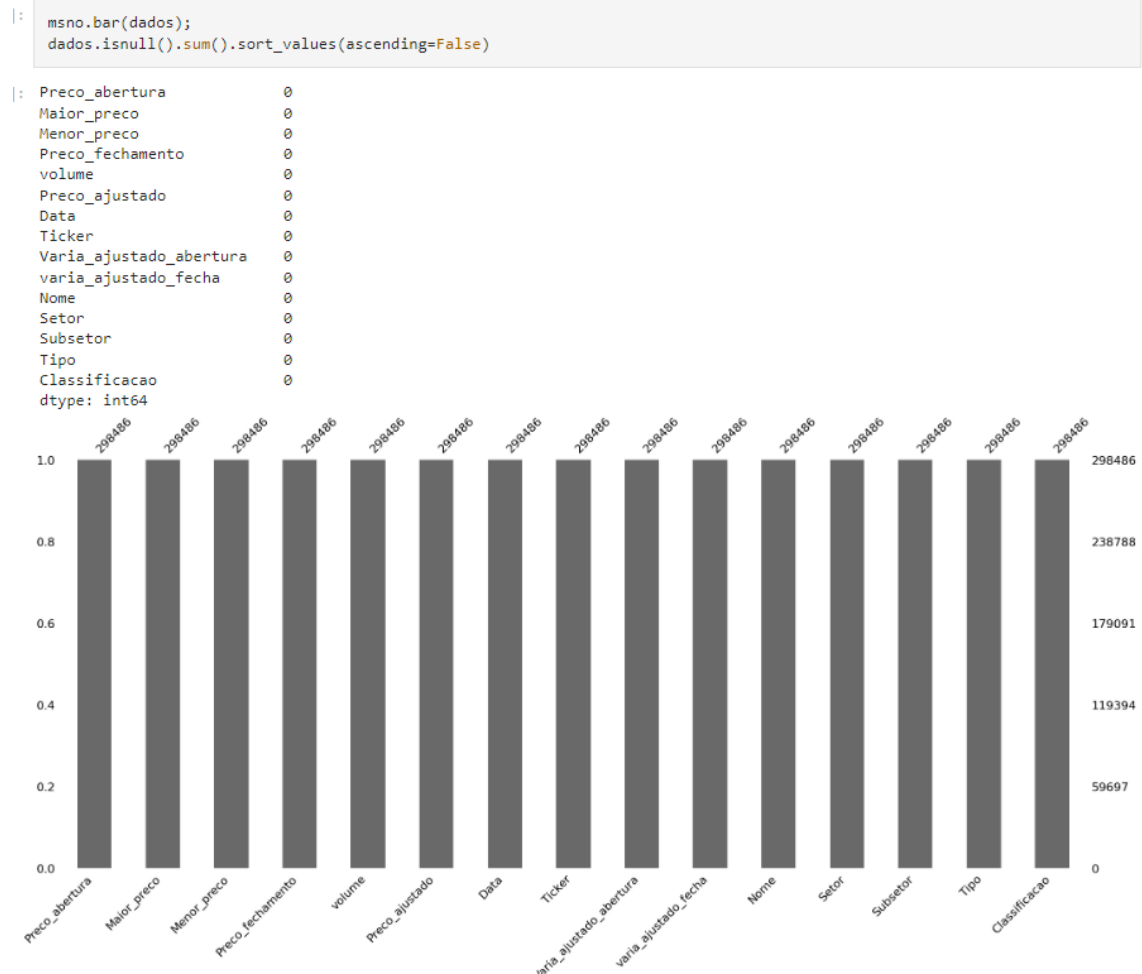


Figura 4 Segunda Verificação

## Carregar dados no serviço do AWS

Após baixar os dados do site da Kaggle e realizar um tratamento inicial dos dados, esse foram carregados na plataforma da AWS.

Inicialmente foi criado uma conta na plataforma da AWS.

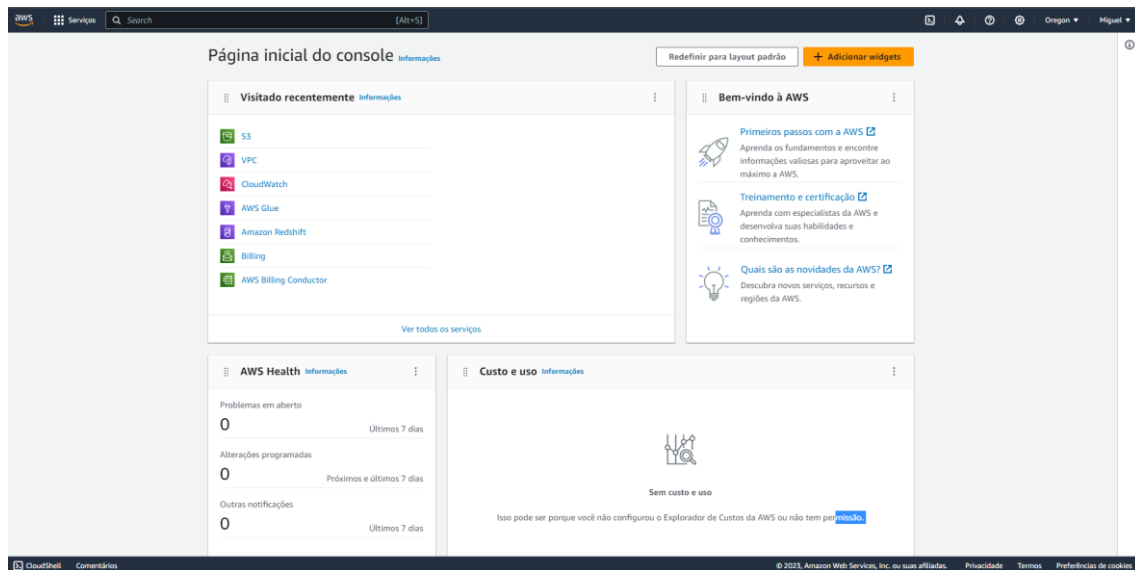


Figura 5 Conta criada na plataforma da AWS

O passo seguinte foi a selecionar ferramenta S3 e cria uma Bucket onde serão carregados o dataset que foi tratado anteriormente.

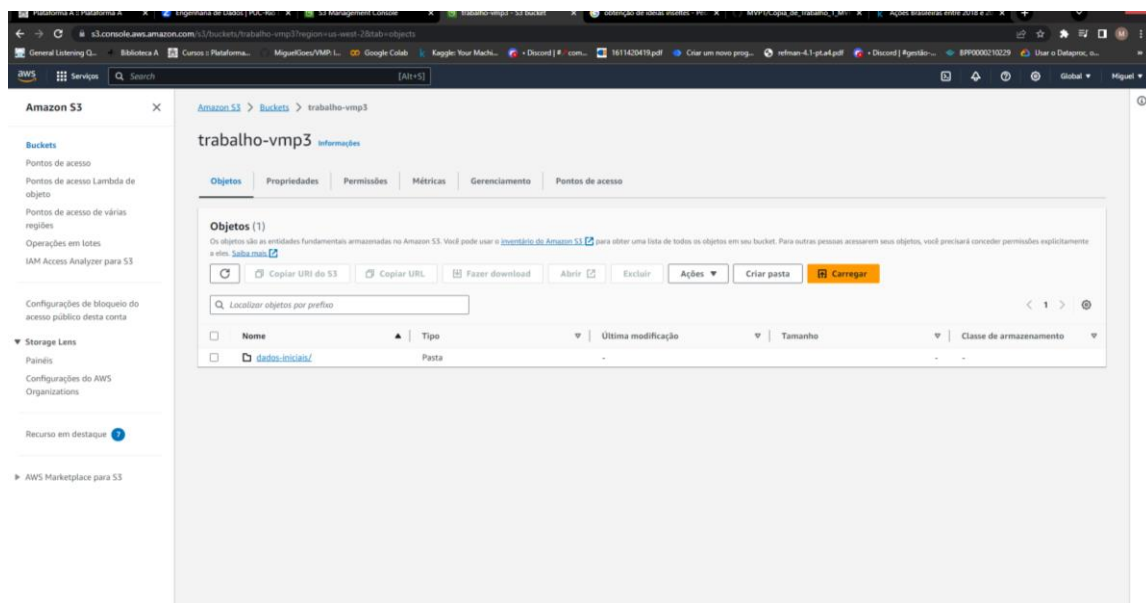


Figura 6 Bucket para o carregamento dos dados

Criado o bucket, foi carregado o arquivo onde se encontra do dataset, nesse caso e possível verificar o tamanho do arquivo e o caminho que é salvo o arquivo

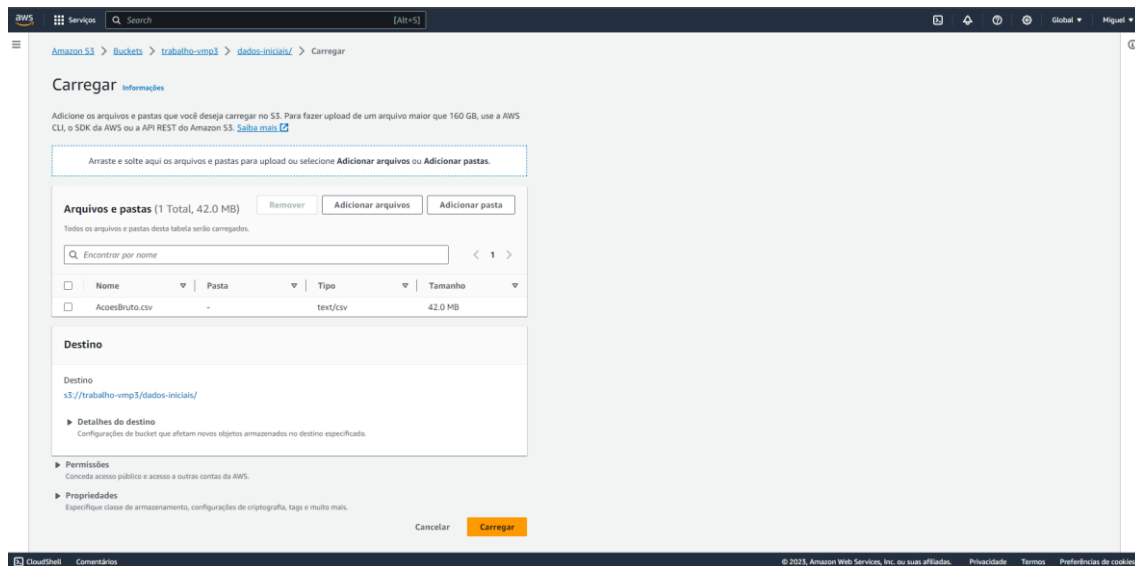


Figura 7 carregamento do arquivo

## Criando um job no AWS GLUE

O passo seguinte é a criação de um job utilizando a ferramenta AWS GLUE. Foi escolhido a configuração da seguinte forma, a fonte dos dados para trabalho foi um bucket do S3 e o target foi o Amazon Redshift, como e mostrado na imagem abaixo.

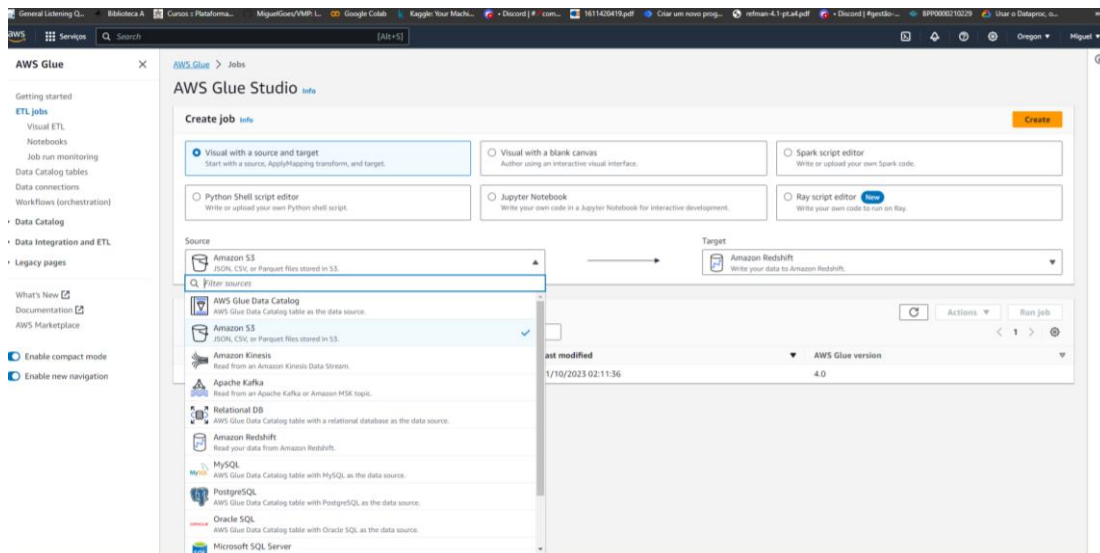


Figura 8 Fonte e destino dos dados

Na aba visualização foi definido o seguinte esquema, os dados são obtidos no bucket do S3 que estão os dados carregados, esses dados sofrem uma transformação utilizando o change schema e por fim são enviados ao Redshift, como indicado na imagem abaixo.

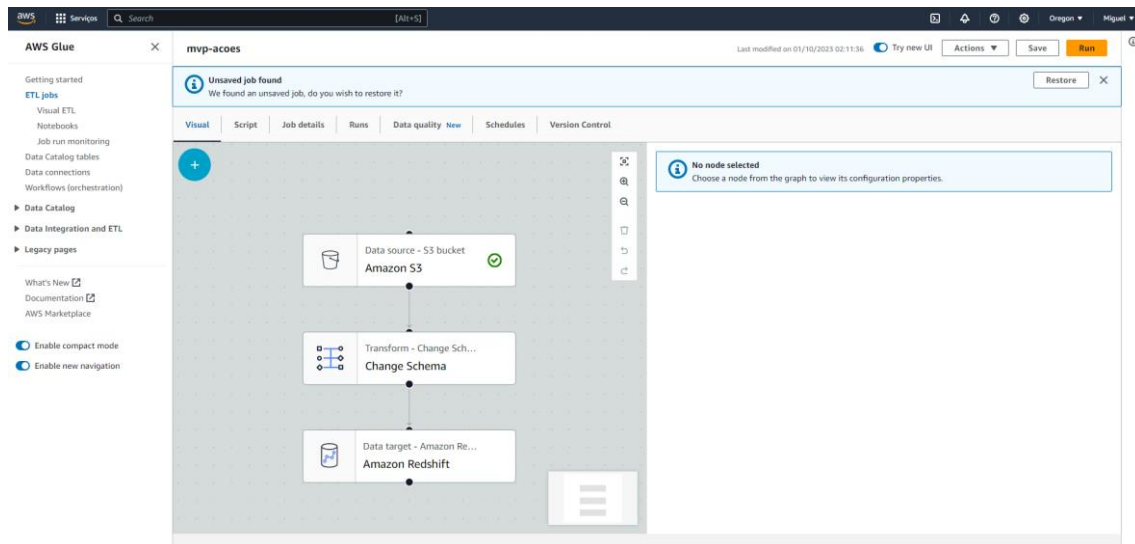


Figura 9 Esquema na aba visual do AWS GLUE

Na primeira parte S3 bucket, foi indicado onde estão carregados os dados na nuvem.

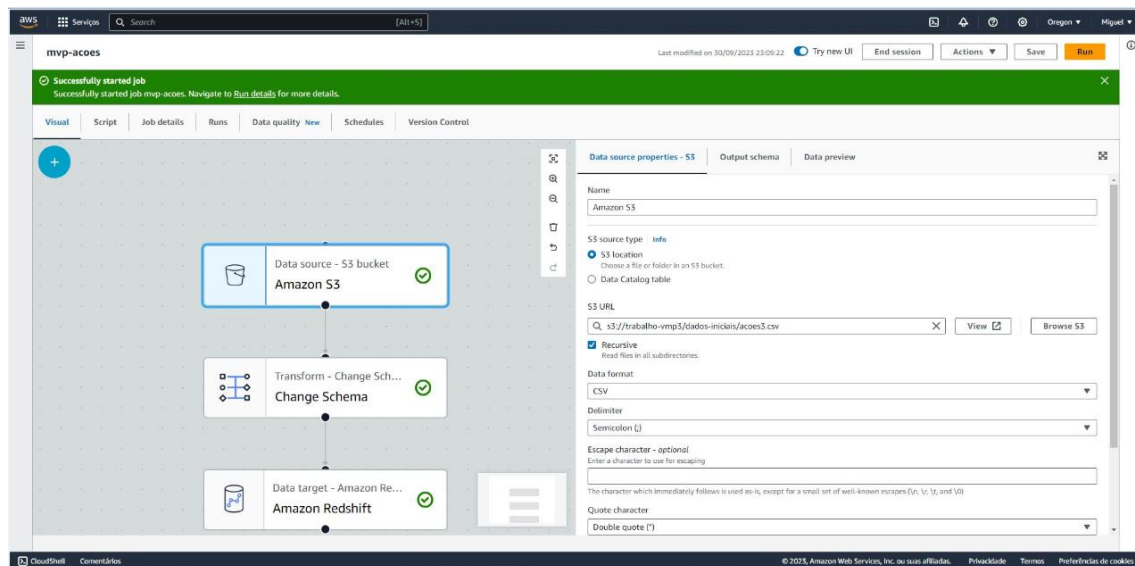


Figura 10 Indicação do local dos dados

No passo seguinte não indicada os dados que podem ser transformados.

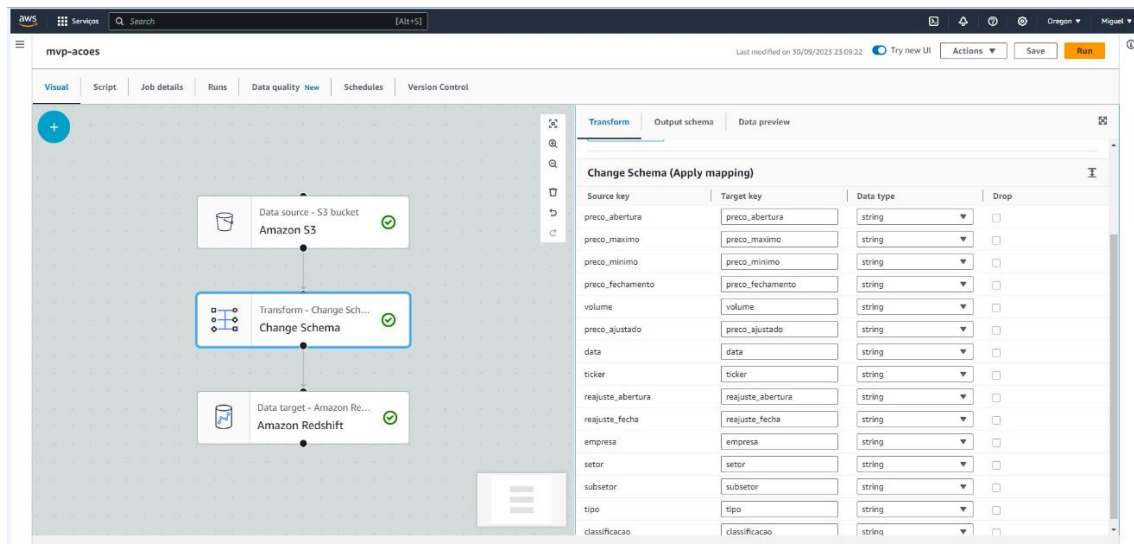


Figura 11 Configuração do Schema

Como os dados que foram carregados possuem muitas informações, e nem todas serão utilizadas, foi feito o drop de algumas colunas deixando apenas a de interesse ativas e alterado o tipo do dado, como mostrado abaixo.

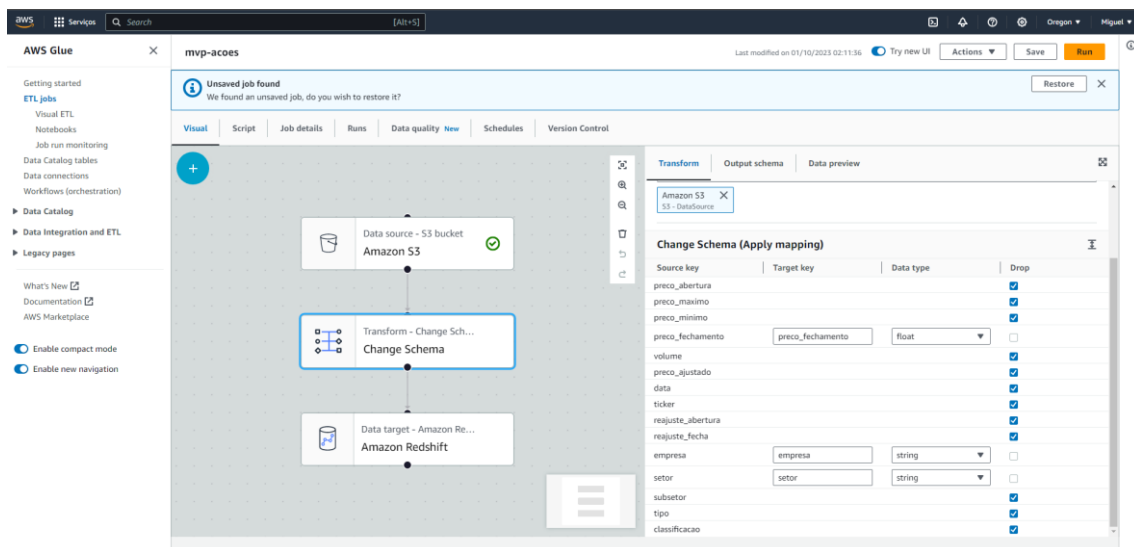


Figura 12 Drop de colunas e alteração do tipo dos dados

Após a transformação é possível conferir quais dados serão enviados ao Redshift, para isso basta clicar na aba "output schema" e a saída é mostrada com na imagem abaixo.



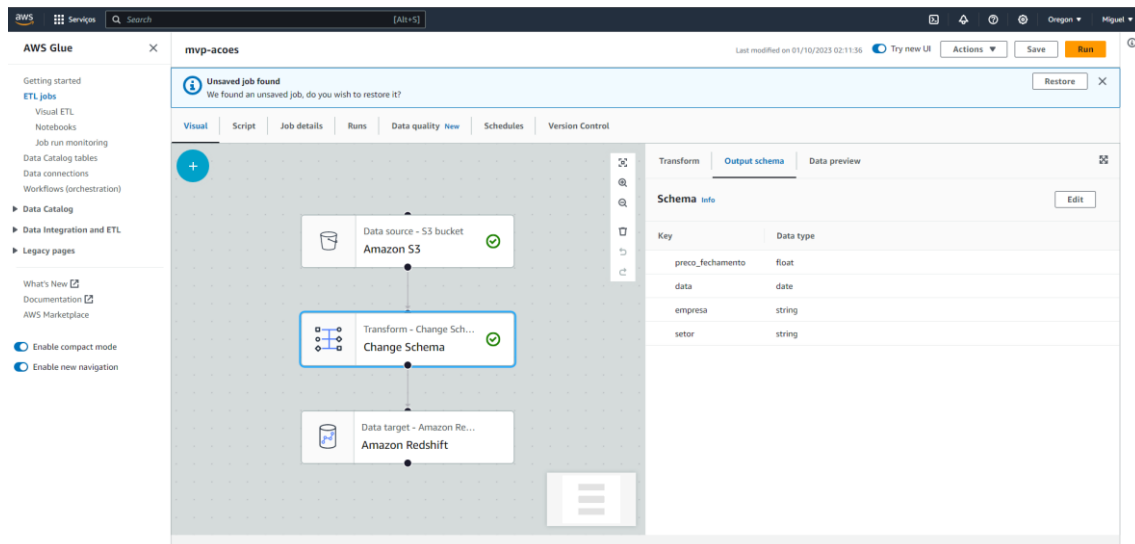


Figura 13 Saída do Schema

## Criação do Redshift

A próxima parte é inserir os dados no Amazon Redshift, para isso será criada uma Amazon Redshift para poder enviar os dados de saída da transformação.

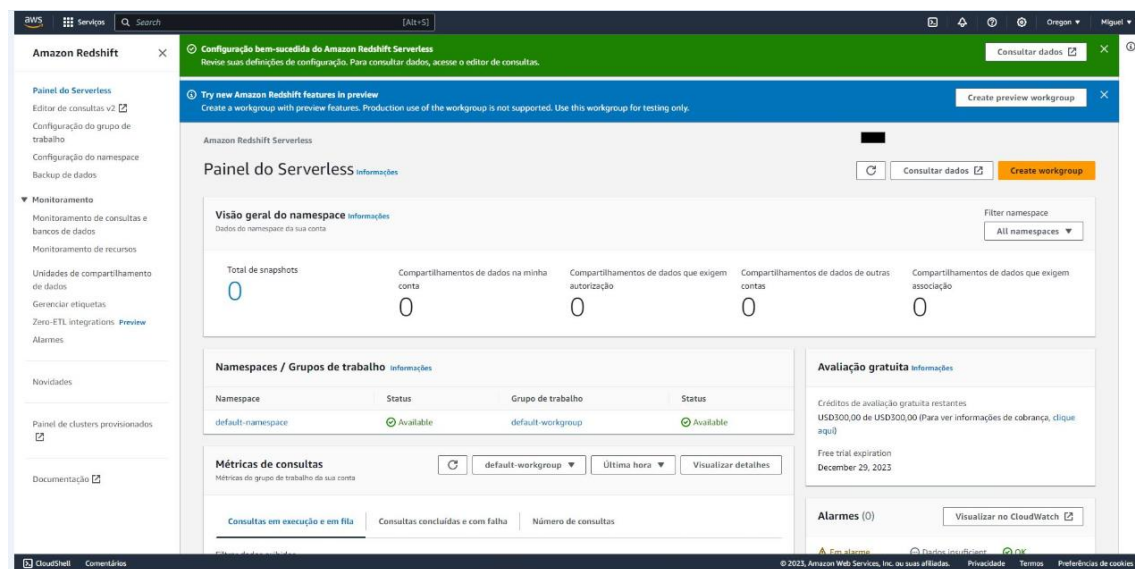


Figura 14 Criação de um redshift

Foram realizadas algumas configuração no redshift para conseguir utilizar a ferramenta conforme foi indicado durante as aulas.

**Capacity**  
Set the base capacity used to process your data warehouse workloads. The capacity is measured in Redshift processing units (RPU). To improve query performance, increase the RPU value.

**Base capacity**  
Base RPU capacity is set to 128 RPUs by default. To change the base RPU capacity, choose another value from the list.

8  
Range must be 8-512 in increments of 8.

**Network and security**

**Virtual private cloud (VPC)**  
This VPC defines the virtual networking environment for this database.

vpc-d9ada2c8b7089d05

**VPC security groups**  
This VPC security group defines which subnets and IP ranges can be used in the VPC.

Choose one or more security groups

sg-0e57a20f2fc9c7f03 X

**Subnet**  
The subnet in the chosen VPC that is associated with the specified database.

Choose three or more subnet IDs

subnet-0c312fcd02fbd3da X subnet-097e42edc23e5638 X  
subnet-005d303921de09697 X subnet-0e9c8178ae410e101 X

**Enhanced VPC routing**  
Turning on this option routes network traffic between your serverless database and data repositories through a VPC instead of the internet.

☐ Turn on enhanced VPC routing

Figura 15 Configuração de subnet indicada

Foi realizada a conexão que será utilizada com a vpc durante a execução do job, após a realização do conexão foi feito o teste da conexão para verificar se estava tudo funcionando como previsto.

Após realizado a configuração acima necessário a criação de um roles para permitir a execução da conexão

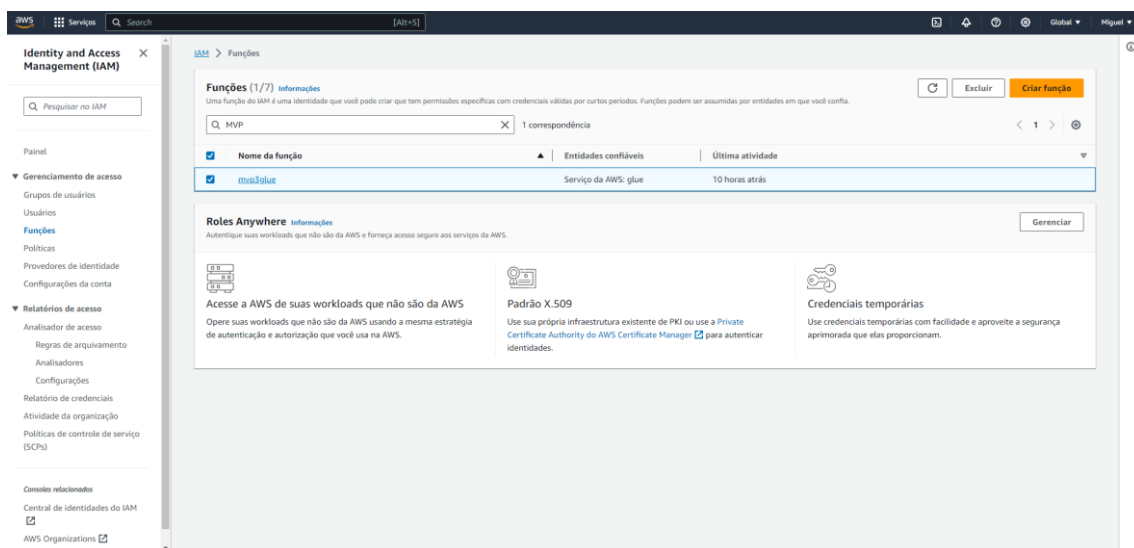
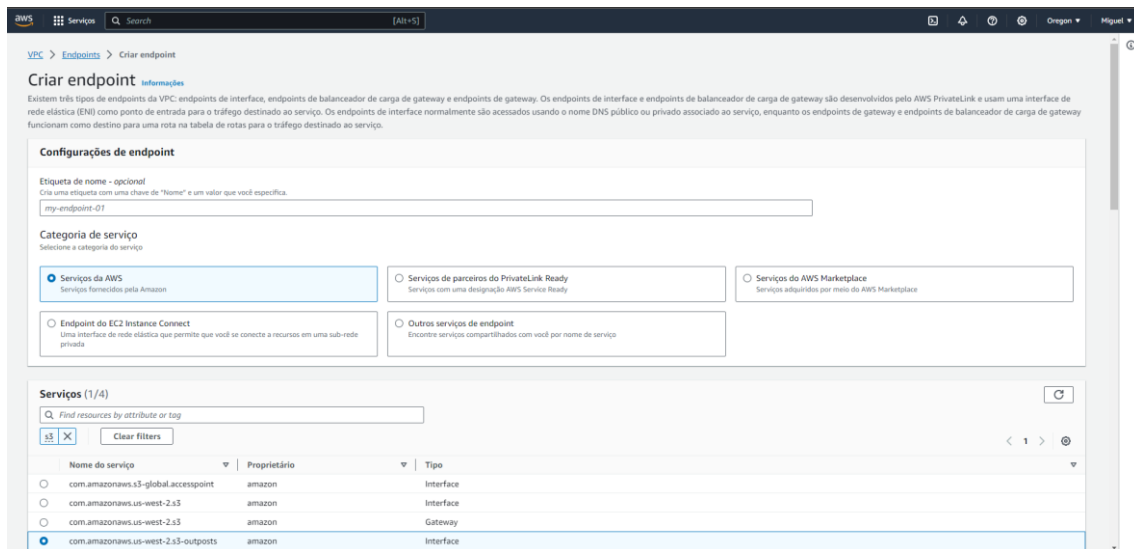


Figura 16 criação de um role

Para permitir que os dados sejam consultados no bucket que está fora da VPC é necessário criar um endpoint, essa criação foi feita como mostrado abaixo.

## Criação do endpoint



Após a criação do endpoint foi realizada o teste da conexão como mostrado

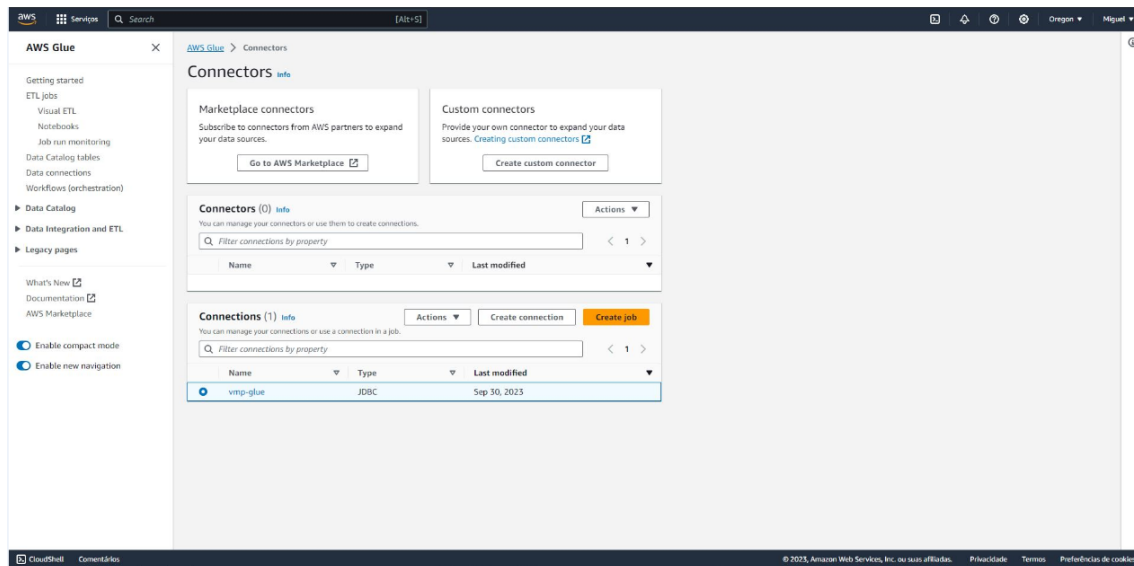


Figura 17 Realização da conexão

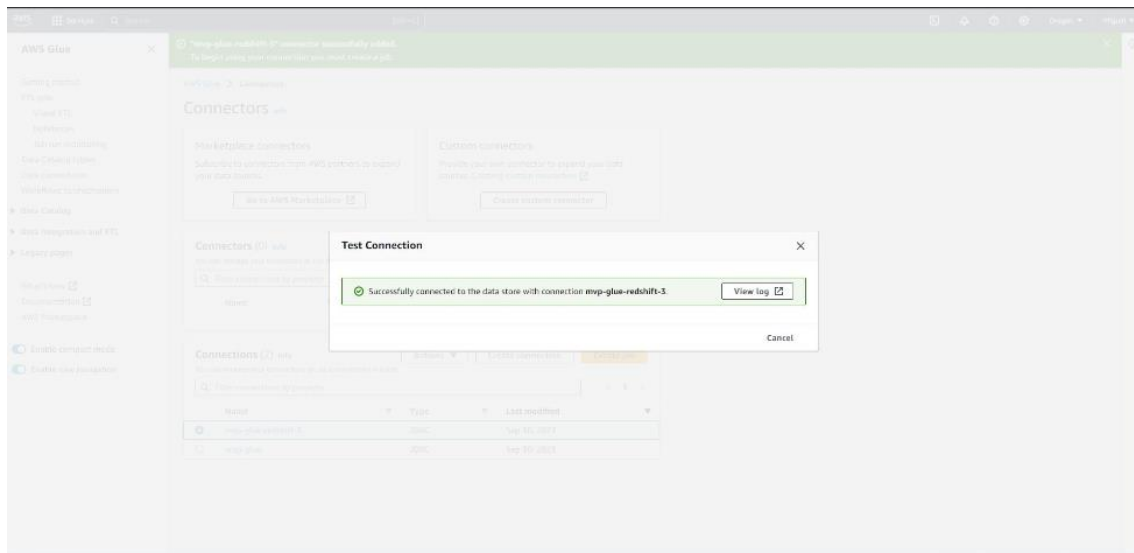


Figura 18 conexão testada com sucesso

A conexão foi realizada com sucesso, isso quer dizer, os dados estão sendo transferidos corretamente do bucket onde estão os dados para dentro do job.

Realizada a conexão, voltei para o job, na parte do Amazon redshift selecionei a conexão que foi criada.

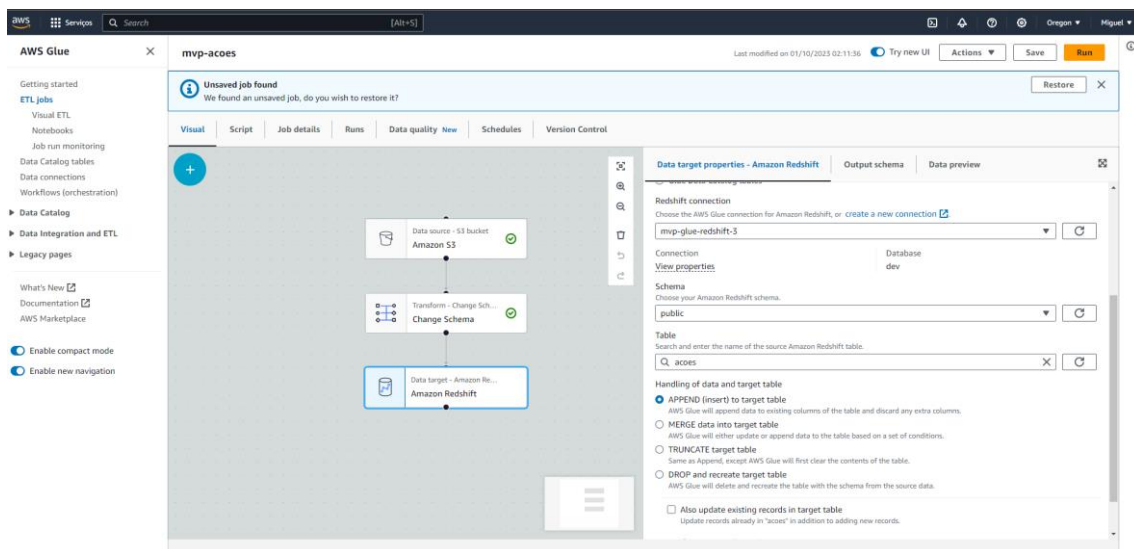


Figura 19 Selecionando a conexão criada

Agora foi criada uma tabela utilizando o Redshift query.

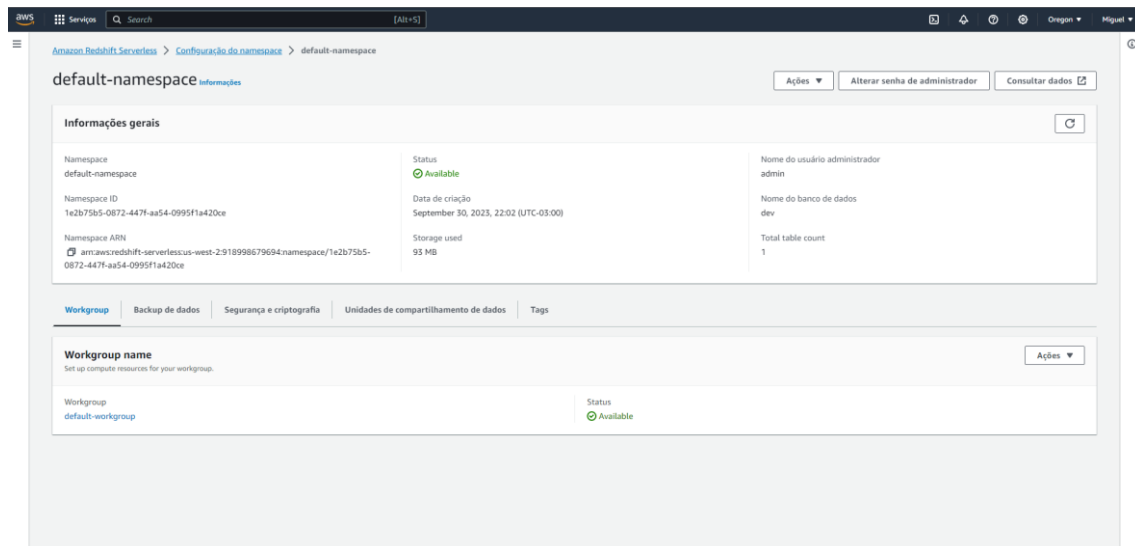


Figura 20 criando uma tabela

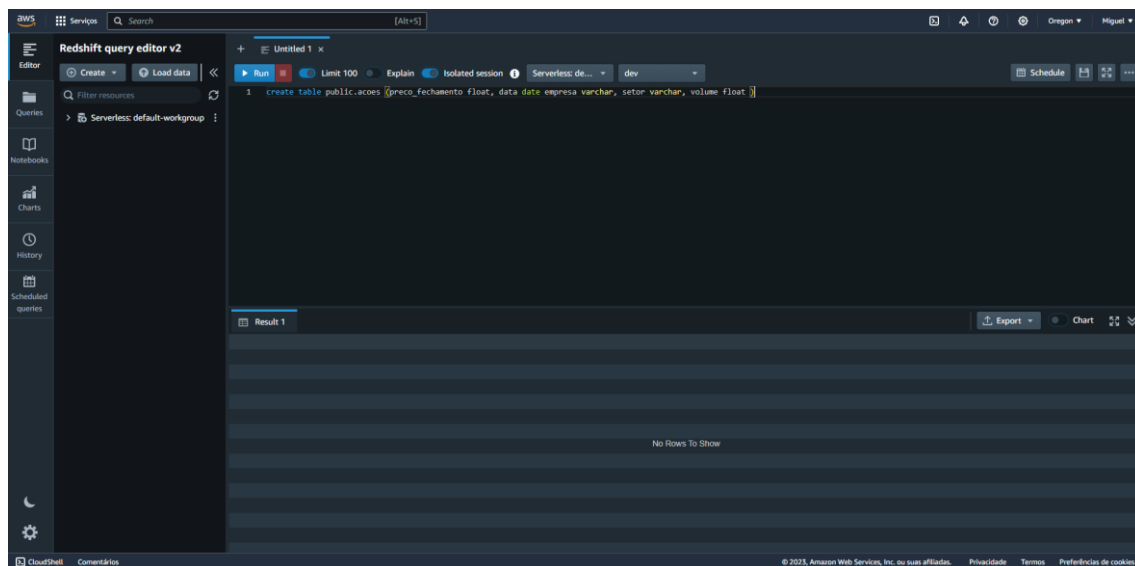


Figura 21 criando uma tabela

Nessa tabela temos o preço de fechamento da ação no dia determinado, o nome da empresa, o volume de ações negociadas e o setor a qual essa empresa pertence.

Ao executar o comando uma tabela é criada com os campos especificados.

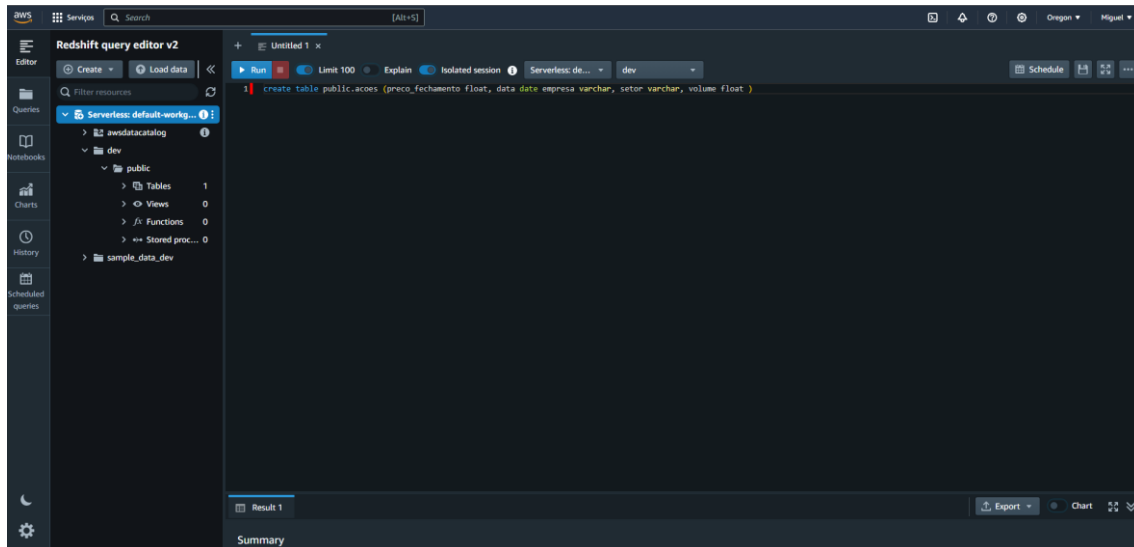


Figura 22 tabela criada

Retornando para a job na parte visual e possível perceber que a tabela foi criada.

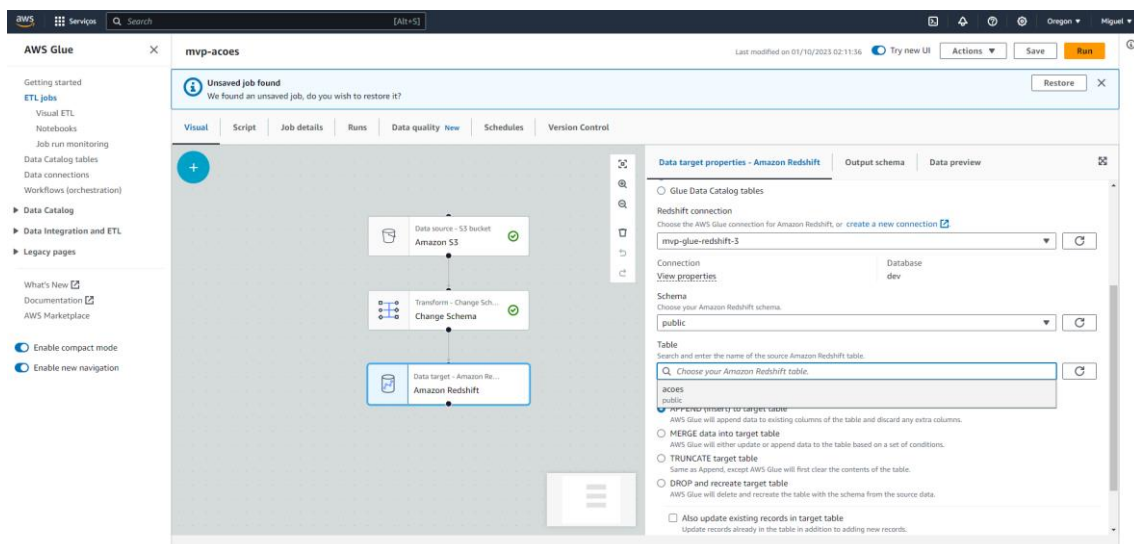


Figura 23 tabela criada e indicada

Foi verificada o datapreview se estava tudo correto.

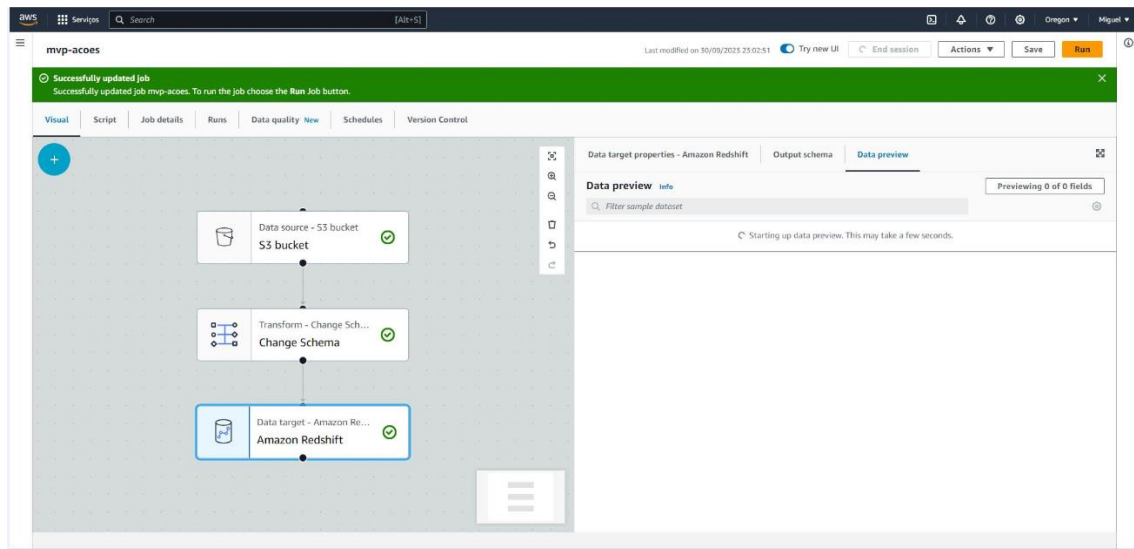


Figura 24 Datapreview executando

Foi feita também a verificação na aba job details dos todos os parâmetros.

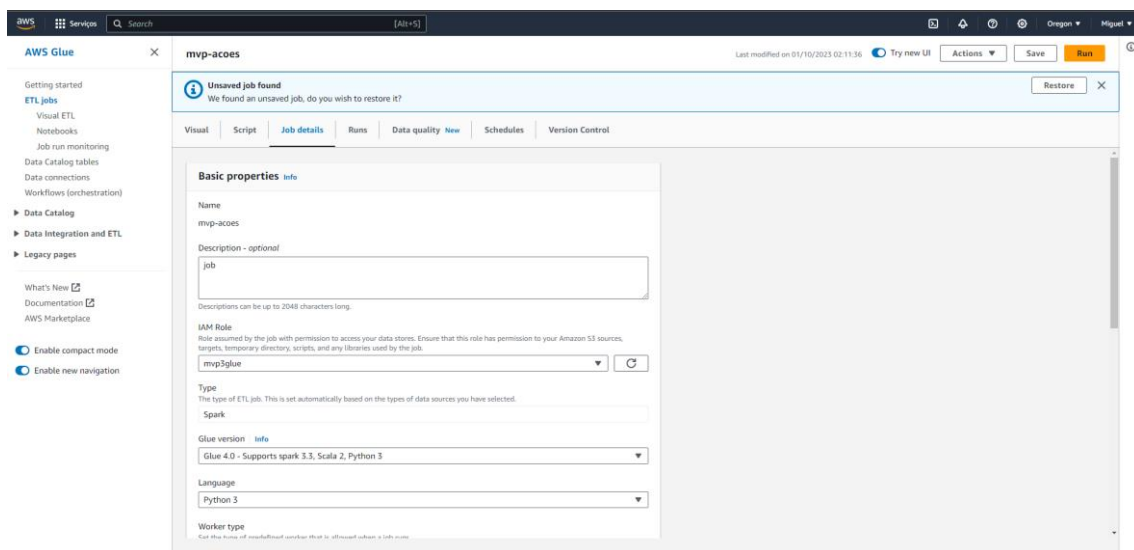


Figura 25 Job details parte 1

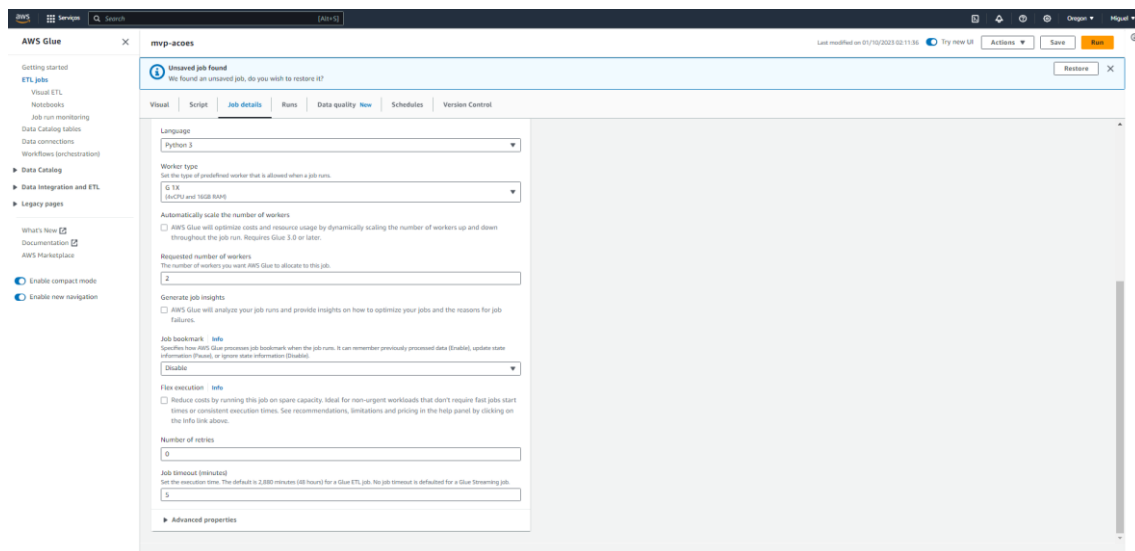
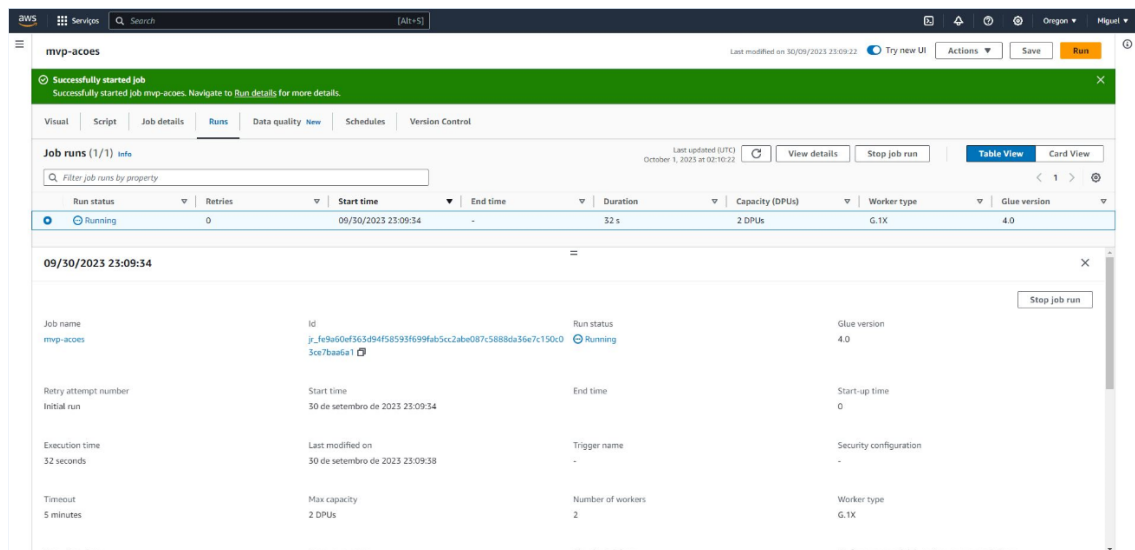


Figura 26 Job details parte2

## Rodando os Job

Após a configuração de todos os pontos e executado o job.



Foram realizados vários testes, os primeiros testes retornaram uma falha, após algumas correções de nomes e do arquivo que foi carregado o job funcionou corretamente.



**Unsaved job found**  
We found an unsaved job, do you wish to restore it?

Visual Script Job details **Runs** Data quality New Schedules Version Control

Job runs (1/8) info  
Last updated: 01/10/2023 02:11:56  
October 1, 2023 at 10:00 AM

Filter job runs by property

Run status	Retries	Start time	End time	Duration	Capacity (DPU)	Worker type	Glue version
Success	0	10/01/2023 01:40:06	10/01/2023 01:41:37	1 m 36 s	2 DPUs	G.1X	4.0
Success	0	10/01/2023 01:09:07	10/01/2023 01:11:17	1 m 54 s	2 DPUs	G.1X	4.0
Success	0	10/01/2023 01:00:23	10/01/2023 01:05:10	2 m 7 s	2 DPUs	G.1X	4.0
Failed	0	10/01/2023 00:53:39	10/01/2023 00:56:43	2 m 47 s	2 DPUs	G.1X	4.0
Failed	0	10/01/2023 00:43:36	10/01/2023 00:46:19	2 m 27 s	2 DPUs	G.1X	4.0
Failed	0	10/01/2023 00:37:53	10/01/2023 00:40:35	2 m 24 s	2 DPUs	G.1X	4.0
Failed	0	10/01/2023 00:05:22	10/01/2023 00:07:02	1 m 22 s	2 DPUs	G.1X	4.0
Failed	0	09/30/2023 23:09:34	09/30/2023 23:11:37	1 m 46 s	2 DPUs	G.1X	4.0

**10/01/2023 01:40:06**

Job name	mvp-acoes	Run status	Success	Glue version	4.0
Job ID	j-35648055da4996d9698291a22f4c00a0euff8a55ecb331f6a118c	Run status	Success	Glue version	4.0
Retry attempt number	Initial run	Start time	01 de outubro de 2023 01:40:06	End time	01 de outubro de 2023 01:41:37
Execution time	1 minute 16 seconds	Last modified on	01 de outubro de 2023 01:41:37	Trigger name	-
Timeout	5 minutes	Max capacity	2 DPUs	Number of workers	2
Execution class	Standard	Log group name	aws-logs-1-67	Cloudwatch logs	View logs
				Performance and debugging recommendations	View recommendations

Figura 27 jobs que foram executados

A principal fonte dos erros que foram encontrados nos primeiros testes indicados acima foram o delimitador de alguns campos da tabela, o que foi corrigido pela modificação direta do arquivo.

Empresas que tiveram os melhores desempenho

Redshift query editor v2

Query: `SELECT DISTINCT empresa, max(preco_fechamento) as maior_preco_fechamento  
FROM sda_tabela  
GROUP BY empresa  
ORDER BY maior_preco_fechamento DESC  
LIMIT 10;`

Result 1

Summary

Por fim foi realizado a consulta na tabela de saída para identificar as melhores empresas.

- 1- WLM IND COM  
WLM atua na produção, criação e comercialização de bovinos de corte, cultivo e comercialização de grãos.
- 2- CEEE-GT  
A Companhia Estadual de Geração e Transmissão de Energia Elétrica (CEEE-GT) é uma empresa de economia mista
- 3- COMGAS  
A Comgás é uma das empresas controladas pela Compass, empresa com objetivo de ampliar e diversificar o mercado de gás natural no Brasil.
- 4- CEB  
A Companhia Energética de Brasília (CEB), controlada pelo Grupo CEB, é uma holding originada da antiga Companhia de Eletricidade de Brasília.
- 5- TEKA  
O Grupo Teka é uma multinacional alemã fundada em 1924, cuja principal atividade é a produção e a comercialização de produtos de cozinha
- 6- B2W DIGITAL  
A B2W Companhia Digital é uma empresa que atua no comércio eletrônico, que surgiu a partir da junção das plataformas digitais Americanas.
- 7- ENCORPAR  
A Encorpar, com sede no estado de Minas Gerais, tem por objetivo a comercialização de fios e tecidos em geral.
- 8- WLM IND COM  
O principal foco de negócios da WLM é a comercialização e a prestação de serviços de manutenção para veículos da marca Scania.
- 9- METALFRIO  
A Metalfrio Solutions é uma empresa global, de origem brasileira, que está entre as líderes mundiais do setor de refrigeração comercial.
- 10- WEG  
Fundada em 1961, a WEG é uma empresa global de equipamentos eletroeletrônicos, que atua no setor de bens de capital com foco em motores