

Curso de Especialização em Big Data – Escola Politécnica da USP

Disciplina de Repositórios de Dados e NoSQL eEDB-016

Prof. Dra. Jeaneth Machicao - Prof. Dr. Pedro Luiz Pizzigatti Corrêa

Projeto Final

Grupo 2

Ingrid Silva

Lucas Pereira

Miguel Ferreira

João Martins

Exercício 1

Passo 1 - Criação de uma instância EC2

Para podermos criar um banco de dados no ambiente da AWS, optamos por criar uma instância EC2 que funcionasse como o recurso necessário para esse banco

Resumo da instância para i-06b228840b219e114 (instancia-ingesao-dados) Informações

Atualizado há less than a minute

ID da instância i-06b228840b219e114	Endereço IPv4 público 54.167.197.142 endereço aberto	Endereços IPv4 privados 172.31.47.179
Endereço IPv6 -	Estado da instância Executando	DNS pública ec2-54-167-197-142.compute-1.amazonaws.com endereço aberto
Tipo de nome do host Nome do IP: ip-172-31-47-179.ec2.internal	Nome do DNS de IP privado (somente IPv4) ip-172-31-47-179.ec2.internal	Endereços IP elásticos -
Nome do DNS do recurso privado de resposta IPv4 (A)	Tipo de instância t3.micro	Descoberta do AWS Compute Optimizer Opte por participar do AWS Compute Optimizer para obter recomendações. Saiba mais
Endereço IP atribuído automaticamente 54.167.197.142 [IP público]	ID da VPC vpc-09163af7241454b9a	Nome do Grupo do Auto Scaling -
Função do IAM -	ID da sub-rede subnet-0040b16602202cc31	Gerenciado falso
IMDSv2 Required	ARN da instância arn:aws:ec2:us-east-1:163080369211:instance/i-06b228840b219e114	
Operador -		

EC2 > Instâncias

Instâncias (1) Informações

[Conectar](#) [Estado da instância](#) [Ações](#) [Executar instâncias](#)

[Todos os estados](#)

<input type="checkbox"/>	Name	ID da instância	Estado da inst...	Tipo de inst...	Verificaç
<input type="checkbox"/>	instancia-inge...	i-06b228840b219e114	Executando	t3.micro	3/3 vi

Passo 2 - Criação do banco de dados no RDS

Optamos por criar o nosso banco de dados utilizando PostgreSQL no RDS, também da AWS

The screenshot displays the AWS Management Console interface for an Amazon RDS database instance. The top section shows the instance details for 'database-ingestao-dados', including its status (Disponível), function (Instância), mechanism (PostgreSQL), and region (us-east-1a). Below this, the 'Segurança e conexão' (Security and connection) tab is selected, showing the endpoint, port, and network configuration. The bottom section shows the 'Bancos de dados (1)' (Databases (1)) list, which includes the instance 'database-ingestao-dados' with a status of 'Disponível' (Available).

database-ingestao-dados

Resumo

Identificador de banco de dados	Status	Função	Mecanismo	Recomendações
database-ingestao-dados	Disponível	Instância	PostgreSQL	2 Informativa
CPU	Classe	Atividade atual	Região e AZ	
3.91%	db.t4g.micro	0.00 sessões	us-east-1a	

Segurança e conexão

Endpoint e porta	Redes	Segurança
Endpoint database-ingestao-dados.ctfjpcnekhp.us-east-1.rds.amazonaws.com	Zona de disponibilidade us-east-1a	Grupos de segurança da VPC rds-ec2-2 (sg-08f4d551db61bd7f9) Ativo
Porta 5432	VPC vpc-09163af7241454b9a	Publicamente acessível Não
	Grupo de sub-redes rds-ec2-db-subnet-group-1	Autoridade de certificação rds-ca-rsa2048-g1
	Sub-redes subnet-04196c0b9dcb098b8 subnet-0e65852d045251720 subnet-09b4ea1cb45218685	Data da autoridade de certificado May 25, 2061, 20:34 (UTC-03:00)

Bancos de dados (1)

Identificador de banco de dados	Status	Função	Mecanismo	Região	Tamanho
database-ingestao-dados	Disponível	Instância	PostgreSQL	us-east-1a	db.t4g.micro

Passo 3 - Armazenamento dos dados no data lake (S3)

Consideramos como ponto de partida da ingestão, os arquivos fornecidos em aula armazenados no data lake da AWS, o S3. Organizamos os temas dos arquivos por pastas.

aws Search [Alt+S] Estados Unidos (Norte da) voclabs/user4245266=Ingrid_Paula_Daniel_Silva @ 1

Amazon S3 > Buckets > eedb-011-ingestao > exercicio-1/

exercicio-1/

Copiar URI do S3

Objetos | Propriedades

Objetos (3)

Copiar URI do S3 Copiar URL Fazer download Abrir Excluir Ações Criar pasta

Carregar

Os objetos são as entidades fundamentais armazenadas no Amazon S3. Você pode usar o [inventário do Amazon S3](#) para obter uma lista de todos os objetos em seu bucket. Para outras pessoas acessarem seus objetos, você precisará conceder permissões explicitamente a eles. [Saiba mais](#)

Localizar objetos por prefixo

<input type="checkbox"/>	Nome	Tipo	Última modificação	Tam
<input type="checkbox"/>	Bancos/	Pasta	-	
<input type="checkbox"/>	Empregados/	Pasta	-	
<input type="checkbox"/>	Reclamações/	Pasta	-	

aws Search [Alt+S] Estados Unidos (Norte d) voclabs/user4245266=Ingrid_Paula_Daniel_Silva @

Amazon S3 > Buckets > eedb-011-ingestao > exercicio-1/ > Reclamações/

Reclamações/

Copiar URI do S3

Objetos | Propriedades

Objetos (7)

Copiar URI do S3 Copiar URL Fazer download Abrir Excluir Ações

Criar pasta Carregar

Os objetos são as entidades fundamentais armazenadas no Amazon S3. Você pode usar o [inventário do Amazon S3](#) para obter uma lista de todos os objetos em seu bucket. Para outras pessoas acessarem seus objetos, você precisará conceder permissões explicitamente a eles. [Saiba mais](#)

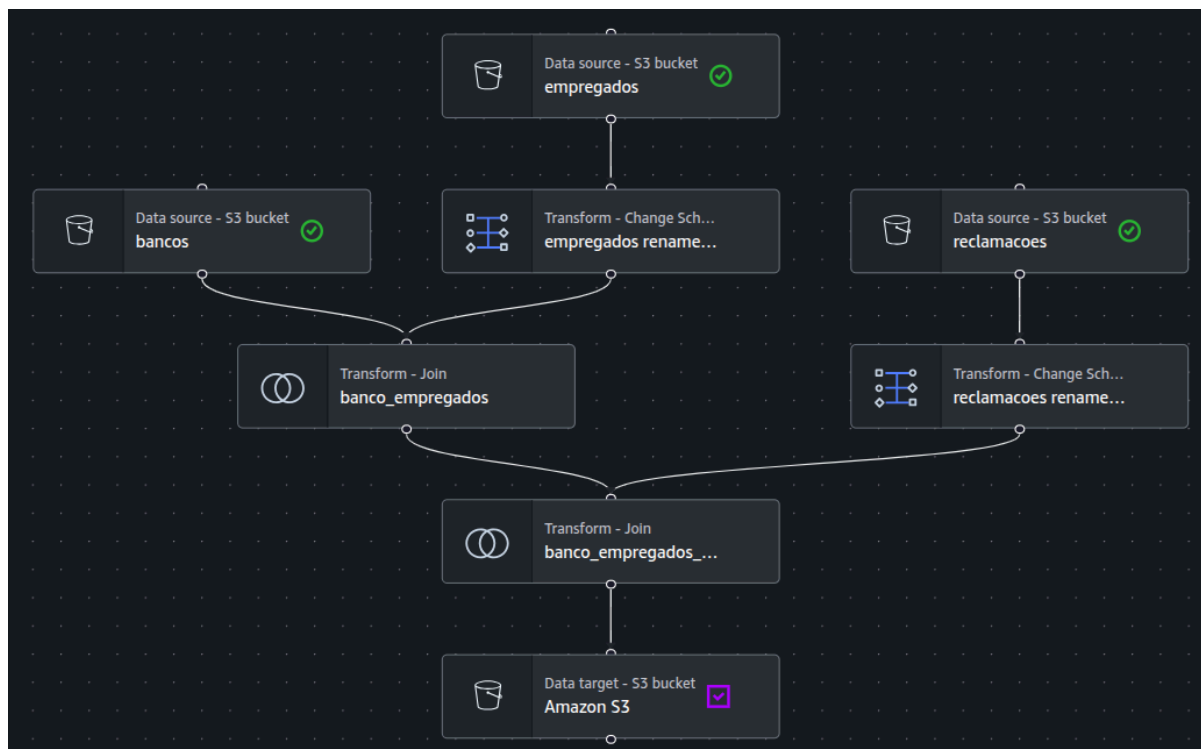
Localizar objetos por prefixo

<input type="checkbox"/>	Nome	Tipo	Última modificação	Tamanho	Classe de armazenamento
<input type="checkbox"/>	2021_tri_01.csv	csv	27 Jul 2025 02:39:24 PM -03	11.3 KB	Padrão
<input type="checkbox"/>	2021_tri_02.csv	csv	27 Jul 2025 02:39:24 PM -03	13.3 KB	Padrão
<input type="checkbox"/>	2021_tri_03.csv	csv	27 Jul 2025 02:39:25 PM -03	13.7 KB	Padrão
<input type="checkbox"/>	2021_tri_04.csv	csv	27 Jul 2025 02:39:25 PM -03	19.9 KB	Padrão
<input type="checkbox"/>	2022_tri_01.csv	csv	27 Jul 2025 02:39:25 PM -03	20.2 KB	Padrão

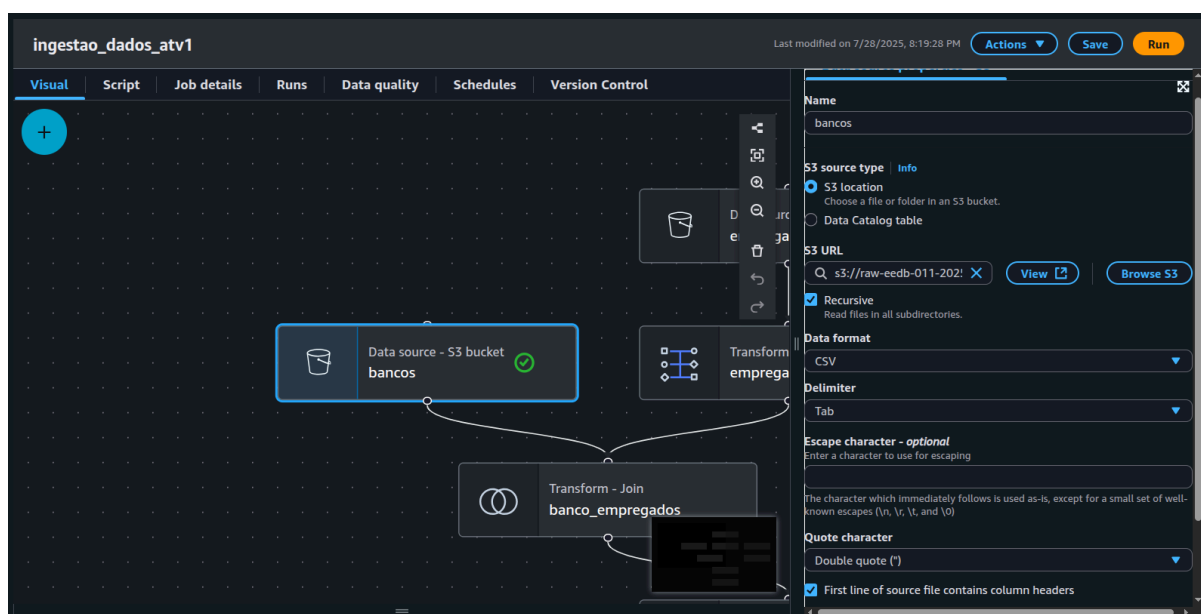
Passo 4 - Tratamento e junção dos dados utilizando ETL Visual (Glue)

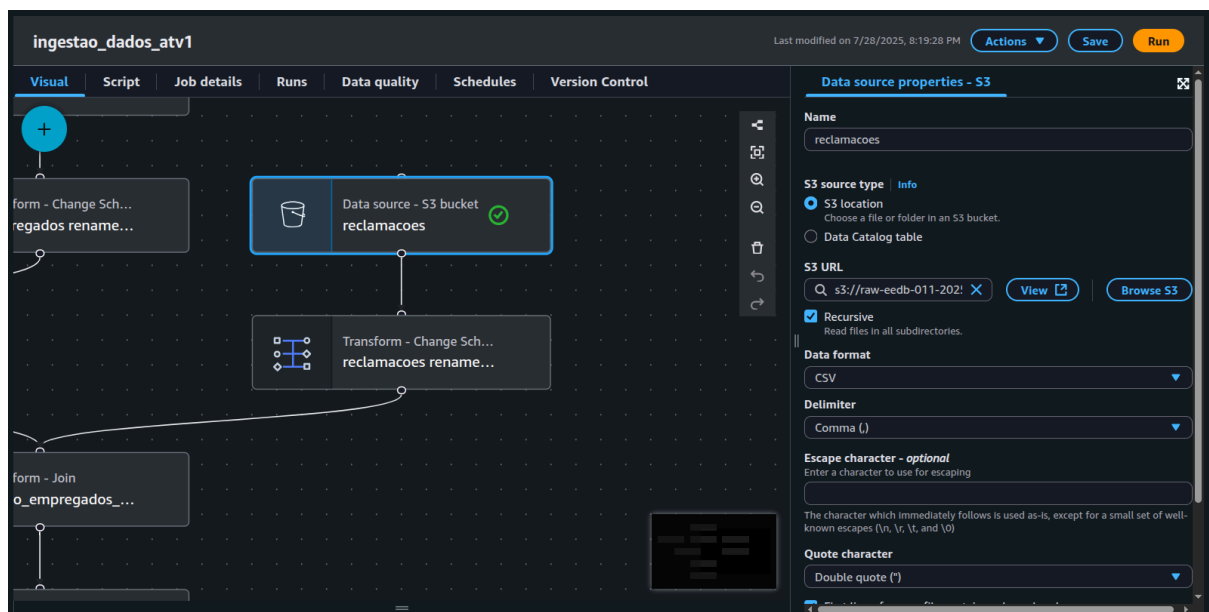
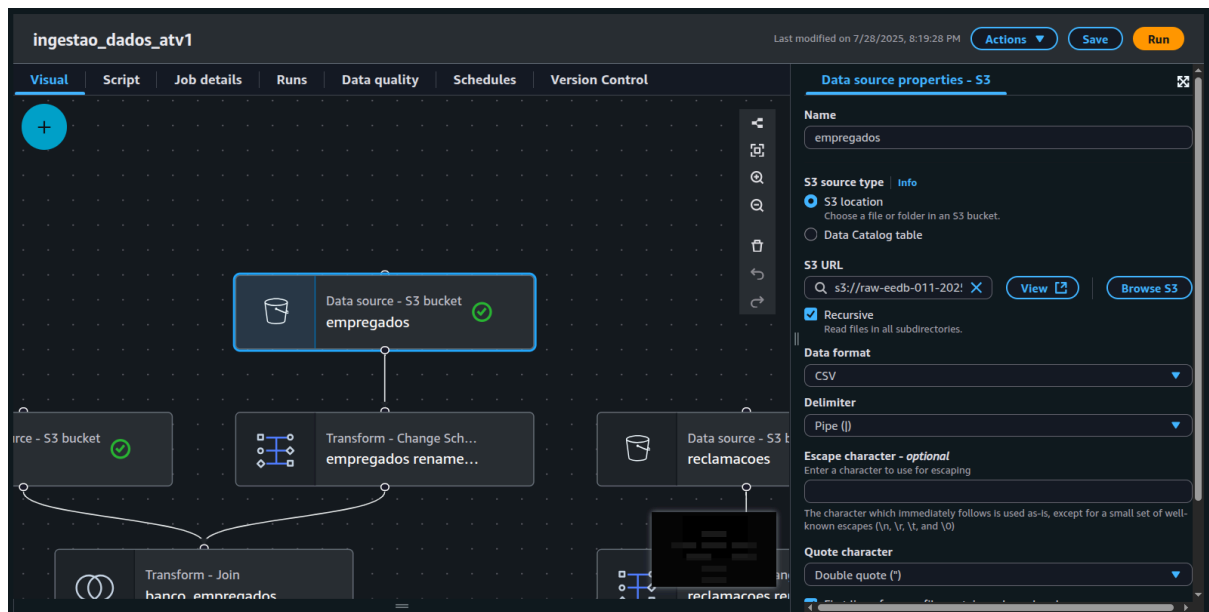
(a) Lendo as três tabelas diretamente do S3

Inicialmente, realizamos a leitura dos dados das três tabelas diretamente do S3. Posteriormente, foram realizadas algumas operações de limpeza e transformações dos dados. E finalmente a união dos dados.



Parâmetros utilizados na leitura.





Resultado

Ao tentar rodar o job ocorreu um erro para que impossibilitou a leitura dos dados de reclamação:

- Unable to parse file: 2022_tri_04.csv

(b) Lendo as duas tabelas diretamente do S3

Dados que existem limites para quantos aos parâmetros que podemos utilizar na leitura desses dados utilizando apenas a interface visual. Não conseguimos prosseguir com a leitura das três tabelas. Sendo assim, nesse momento realizamos a tarefa com apenas o join entre as tabelas de bancos e empregados.



Resultado

Como é possível observar na imagem a seguir, o job funcionou corretamente. Na seção seguinte será possível observar os dados resultantes.

The screenshot shows the AWS Glue console interface for the job 'ingestao_dados_atv1-copy'. The job is in the 'Runs' tab, showing a table of job runs. The job run is successful, with a status of 'Succeeded'. The table below shows the job details.

Run status	Retries	Start time (Local)	End time (Local)	Duration	Capacity (DPUs)	Worker type	Glue version
Succeeded	0	07/28/2025 20:17:57	07/28/2025 20:19:21	1 m 16 s	10 DPUs	G.1X	5.0

Job name	Start time (Local)	Glue version	Last modified on (Local)
ingestao_dados_atv1-copy	07/28/2025 20:17:57	5.0	07/28/2025 20:19:21

Id	End time (Local)	Worker type	Log group name
jr_df08e4286f65cf88252ba1c3a424ba7face9b44a1b	07/28/2025 20:19:21	G.1X	/aws-glue/jobs

Run status	Start-up time	Max capacity	Number of workers
Succeeded	8 seconds	10 DPUs	10

Retry attempt number	Execution time	Execution class	Timeout
-	1 minute 16 seconds	Standard	480 minutes

Initial run	Cloudwatch logs	Usage profile
True	Output logs Error logs	-

Trigger name	Security configuration
-	-

Job run queuing
False

(c) Lendo as três tabelas diretamente do Catálogo

Como uma última tentativa para a leitura dos dados de reclamação, os mesmos foram catalogados utilizando um Crawler, para verificar se assim seria possível realizar a leitura.

Crawler

ingestao_de_dados

Last updated (UTC)
July 30, 2025 at 02:09:33

Run crawler

Edit

Delete

Crawler properties

Name

ingestao_de_dados

IAM role

LabRole

Database

ingestao_de_dados

State

READY

Description

-

Security configuration

-

Lake Formation configuration

-

Table prefix

-

Maximum table threshold

-

Advanced settings

Crawler runs

Schedule

Data sources

Classifiers

Tags

Crawler runs (1)

The list of crawler runs for this crawler.

Stop run

View CloudWatch logs

View run details

Filter data

Filter by a date and time range

< 1 >

Start time (UTC)

End time (UTC)

Current/last duration

Status

DPU hours

Table changes

July 28, 2025 at 23:27:54

July 28, 2025 at 23:29:11

01 min 16 s

Completed

0.045

3 table changes
0 partition changes

Catálogo

Announcing new optimization features for Apache Iceberg tables
Optimize storage for Apache Iceberg tables with automatic snapshot retention and orphan file deletion. [Learn more](#)

ingestao_de_dados

Last updated (UTC)
July 30, 2025 at 02:10:11

Edit

Delete

Database properties

Name

ingestao_de_dados

Description

-

Location

s3://raw-eeedb-011-2025-3-472916995593/resultado/

Created on (UTC)

July 27, 2025 at 19:29:24

Tables (4)

Last updated (UTC)
July 30, 2025 at 02:10:14

Delete

Add tables using crawler

Add table

View and manage all available tables.

Filter tables

< 1 >

Name

Database

Location

Classification

Deprecated

View data

Data quality

Column statistics

atividade1

ingestao_de_dados

s3://raw-eeedb-011-2025-3-472916995593/resultado/atividade1

Parquet

-

Table data

View data quality

View statistics

bancos

ingestao_de_dados

s3://raw-eeedb-011-2025-3-472916995593/resultado/bancos

CSV

-

Table data

View data quality

View statistics

empregados

ingestao_de_dados

s3://raw-eeedb-011-2025-3-472916995593/resultado/empregados

CSV

-

Table data

View data quality

View statistics

reclamacoes

ingestao_de_dados

s3://raw-eeedb-011-2025-3-472916995593/resultado/reclamacoes

CSV

-

Table data

View data quality

View statistics

Athena - É ler os dados de reclamação catalogados através do Athena.

Data

↺

↻

Data source

AwsDataCatalog

Catalog

None

Database

ingestao_de_dados

Tables and views

Create

Filter tables and views

Tables (4)

1

atividade1

⋮

bancos

⋮

empregados

⋮

reclamacoes

⋮

Views (0)

1

Query 1

⋮

1 select * from bancos;

2 select * from empregados;

3 select * from reclamacoes;

4

SQL

Ln 3, Col 1

⋮

ⓘ

⚙

Run again

Explain

Cancel

Clear

Create

Reuse query results

up to 60 minutes ago

Query results

Query stats

Completed

Time in queue: 114 ms

Run time: 843 ms

Data scanned: 125.37 KB

Results (918)

Copy

Download results CSV

Search rows

#

ano

trimestre

categoria

tipo

cnj if

instituio financeira

1

2021

2

Demais bancos e financeiras

Conglomerado

ABC-BRASIL (conglomerado)

2

2021

2

Demais bancos e financeiras

Conglomerado

AGIBANK (conglomerado)

3

2021

2

Demais bancos e financeiras

Banco/financeira

36321990

AGORACRED S/A SOCIEDADE DE CRDITO, FINANCIAMENTO E INVESTIMENTO

4

2021

2

Demais bancos e financeiras

Banco/financeira

27214112

ALS S.A. CRDITO, FINANCIAMENTO E INVESTIMENTO

5

2021

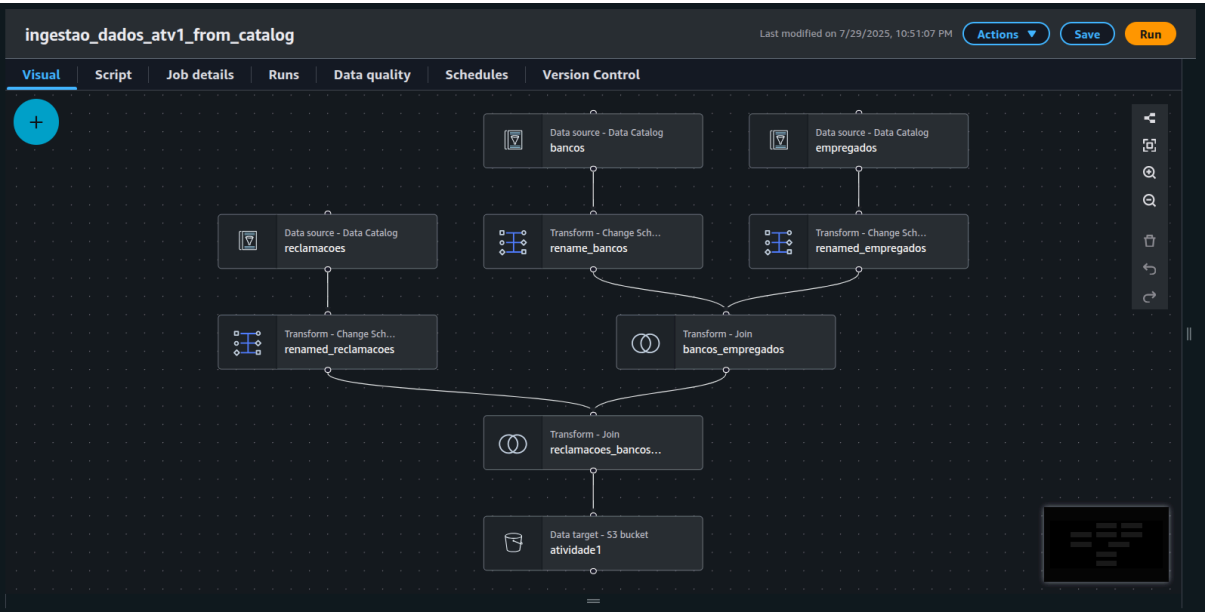
2

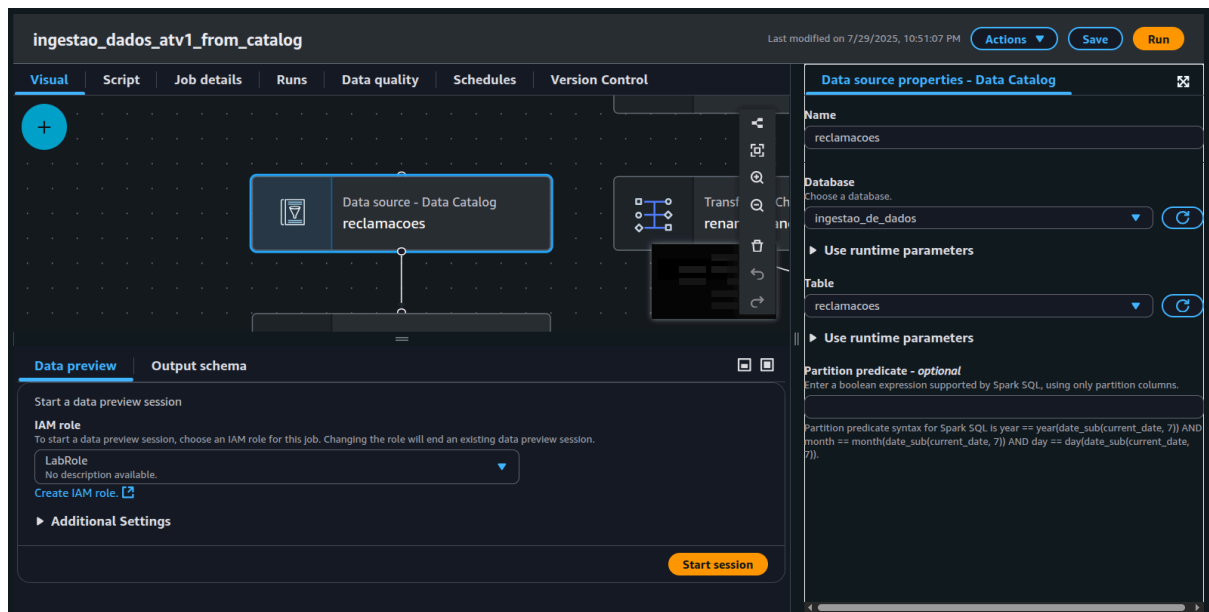
Demais bancos e financeiras

Conglomerado

ALFA (conglomerado)

Job





Resultado

Ao tentar rodar o job ocorreu o mesmo erro para que impossibilitou a leitura dos dados de reclamação:

- Unable to parse file: 2022_tri_01.csv

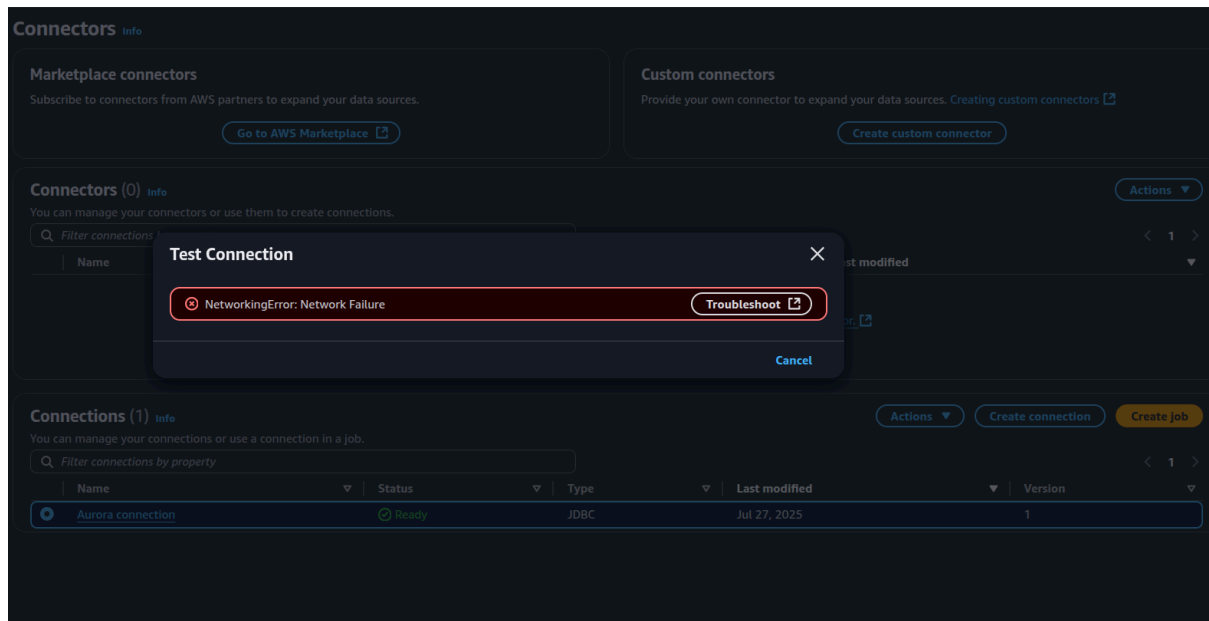
Passo 5 - Ingestão dos dados tratados no banco de dados criado

Como será apresentado na seção extra não conseguimos realizar a conexão com o banco de dados RDS, por conta de questões de configuração de redes. Por conta disso, para ter um resultado enviamos os dados resultantes da execução com sucesso do JOB GLUE para o S3 e também realizamos a catalogação, permitindo assim o acesso aos dados através do Athena.

Extra - Bloqueios e desafios encontrados

Não conseguimos realizar a conexão com o banco de dados RDS, por conta de questões de configuração de redes. Por conta disso, não foi possível realizar a importação dos dados.

(a) Tentativa de conexão pelo Glue



(b) Tentativa de conexão pelo AWS Databrew

aws

Search [Alt+S]

Estados Unidos (Norte d

voclabs/user4245266=Ingrid_Paula_Daniel_Silva @

teste

Conjunto de dados: [exercicio-1-reclamacoes](#)

Projeto: [exercicio-1-ingestao](#)

Receita: [exercicio-1-ingestao-recipe](#)

Executar trabalho

Ações

ABRIR PROJETO

Histórico de execução de trabalhos

Detalhes do trabalho

Linhagem de dados

Receita

Última execução do trabalho **um dia**, nenhuma execução de trabalho agendada

Histórico de execução de trabalhos

Interromper a execução do trabalho

Ações

Pesquisar por ID de execução de trabalho

Mostrar tudo

ID da execução do trabalho	Status da última execução do trabalho	Tempo de execução	Saída	Resumo
teste_2025-07-28-21:43:47	Com falha	Não disponível	1 saída	Account Ver mais
teste_2025-07-28-21:39:35	Com falha	Não disponível	1 saída	The max Ver mais
teste_2025-07-28-21:35:32	Com falha	Não disponível	1 saída	Account Ver mais

Detalhes do erro

AccountId:057688645490 and JobName:163080369211_teste and JobRunId:db_da5efaa03cbf86c57a66537dcd63b285480494009c94b2caec78c4a010801be1 failed to execute with exception Failed to connect to VPC. At least one security group must open all ingress ports. To limit traffic, the source security group in your inbound rule can be restricted to the same security group. Failed to connect to VPC. At least one security group must open all egress ports. To limit traffic, the source security group in your outbound rule can be restricted to the same security group. (Service: AWSGlueDataBrewJobExecutor; Status Code: 400; Error Code: InvalidInputException; Request ID: 4980bc9e-b5d5-44f0-8217-54757c5221ed; Proxy: null)

Fechar