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

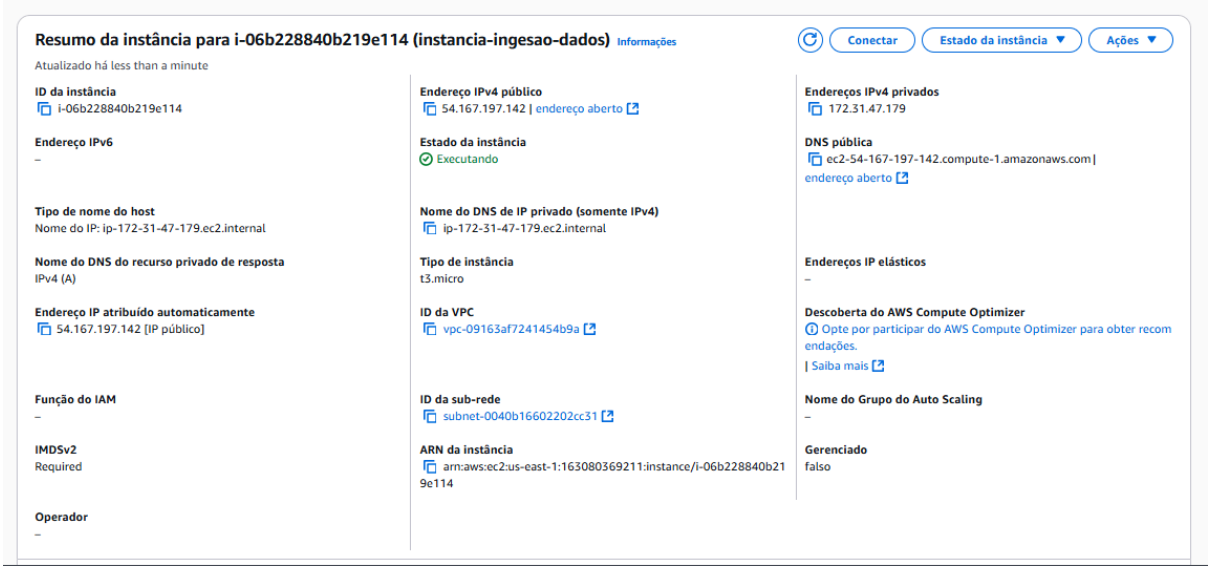
Miguel Ferreira

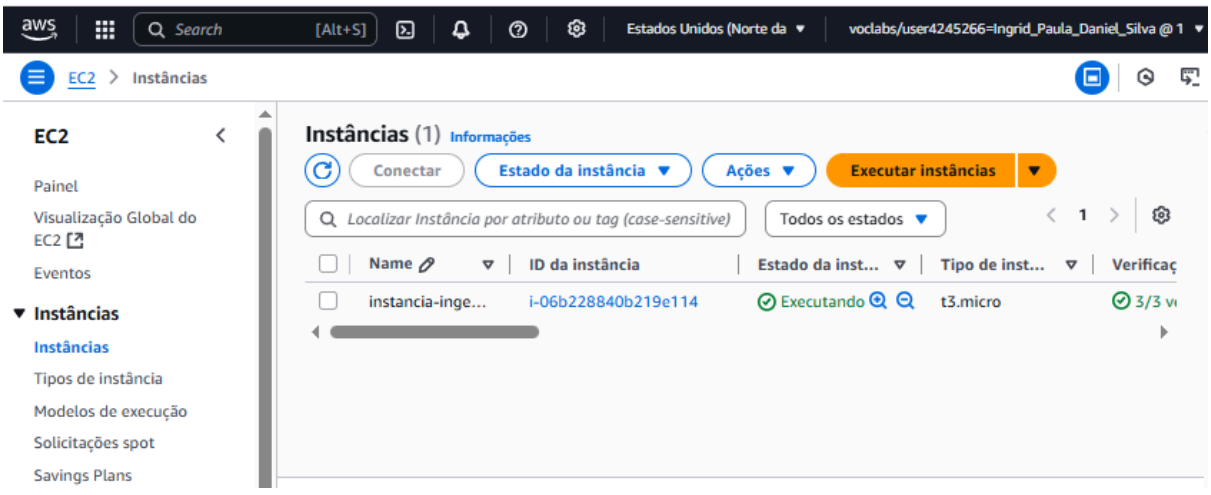
João Martins

Exercício 1

Passo 1 - Criação de uma instância EC2

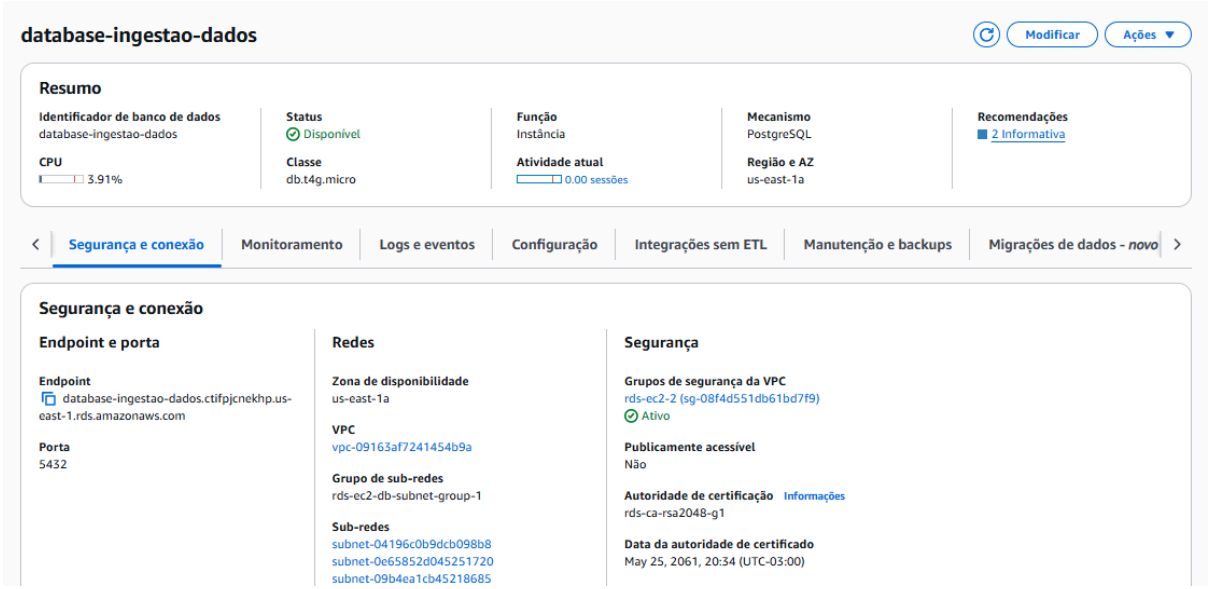
Para podermos criar um banco de dados no ambiente da AWS, optamos por criar uma instância EC2 que funcionasse como o recurso necessário para esse banco

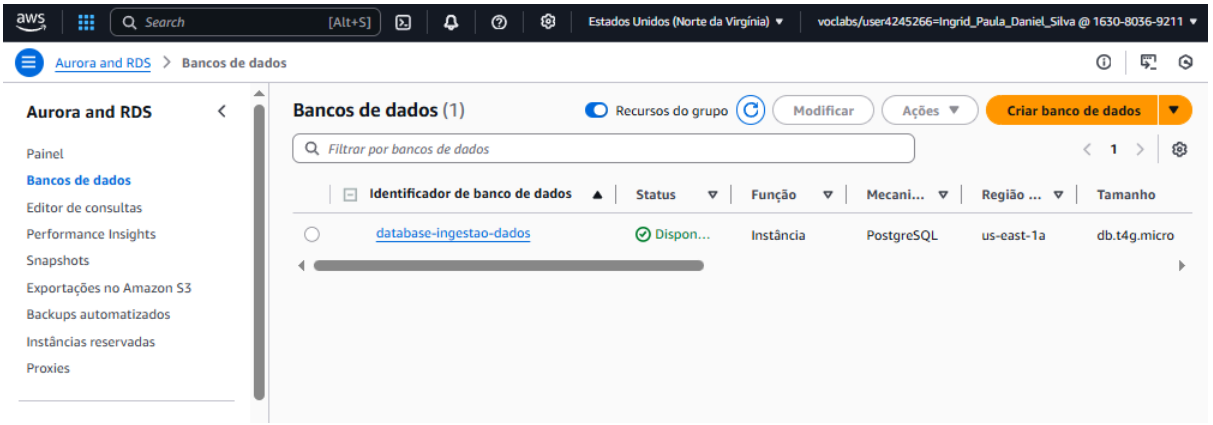




Passo 2 - Criação do banco de dados no RDS

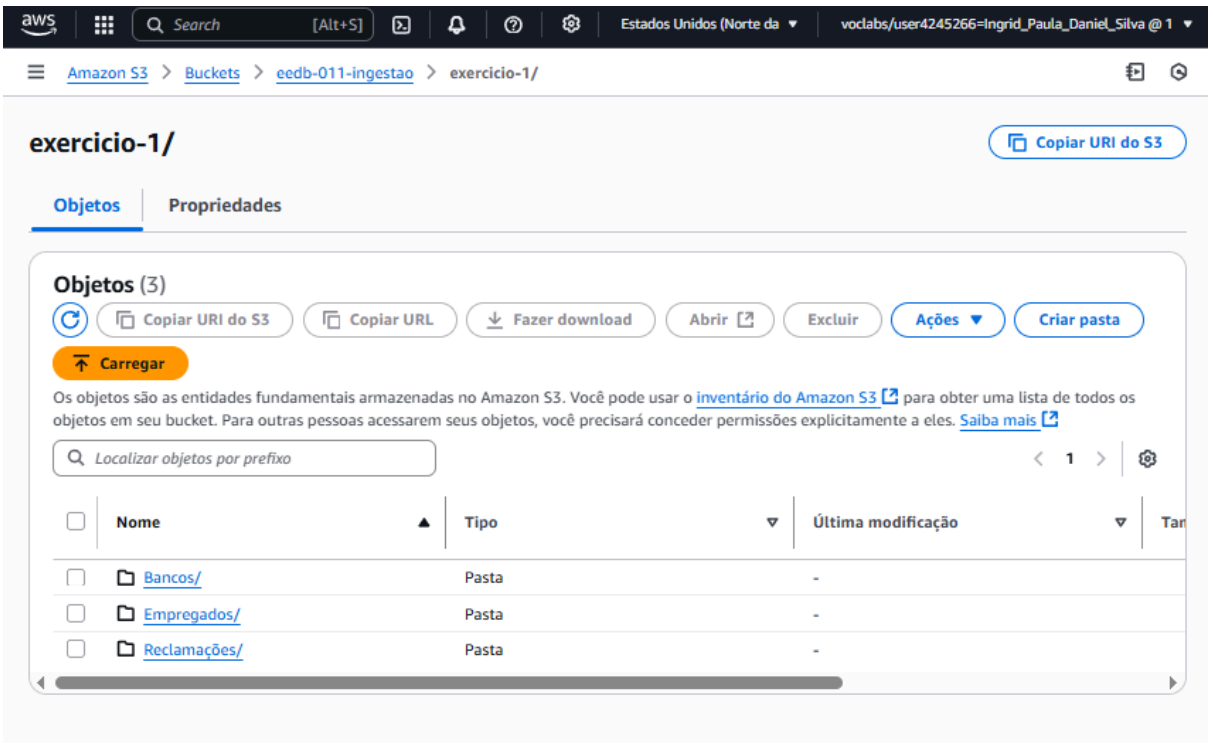
Optamos por criar o nosso banco de dados utilizando PostgreSQL no RDS, também da AWS

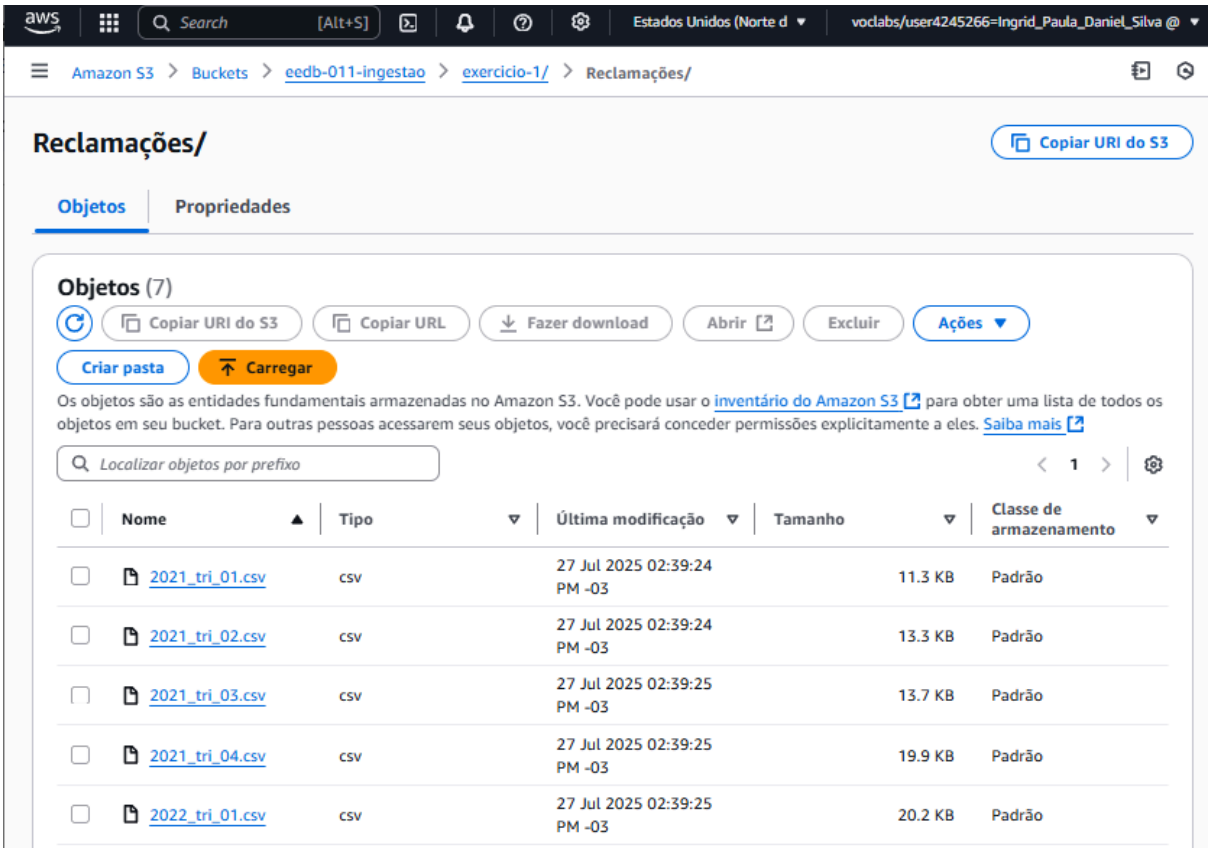




Passo 3 - Armazenamento dos dados no data lake (S3)

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.

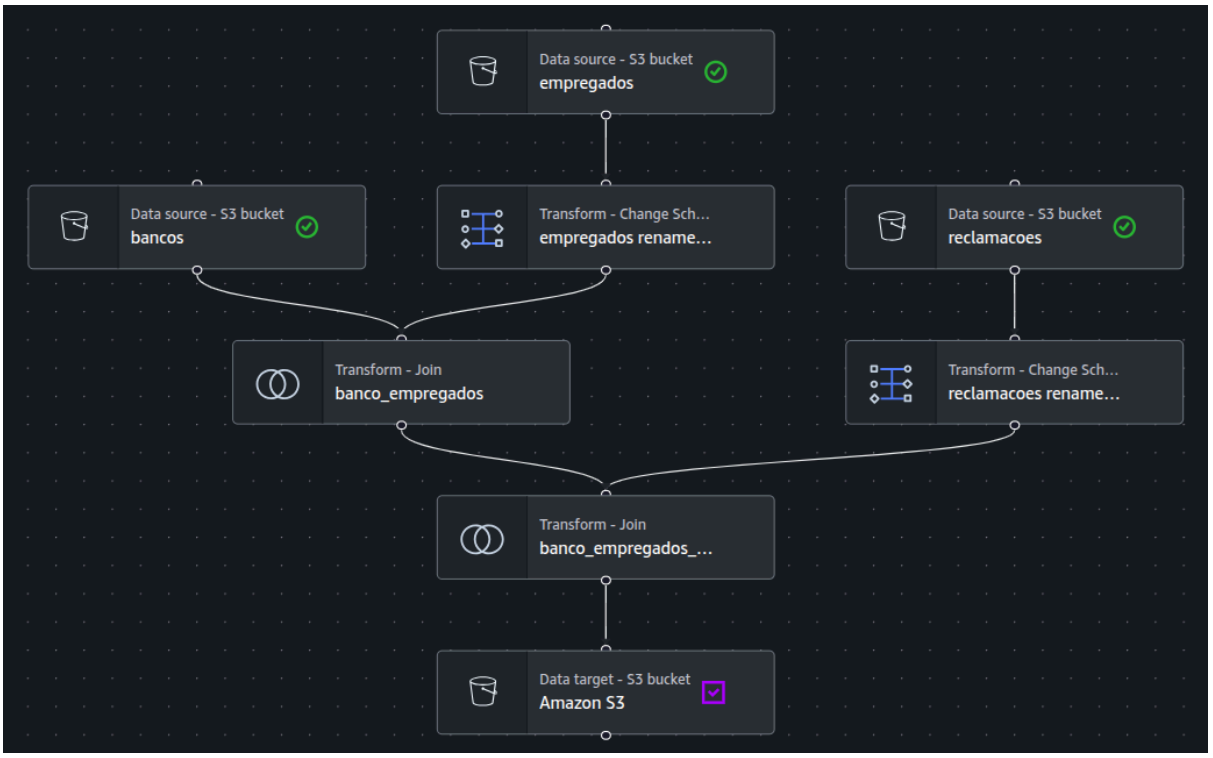




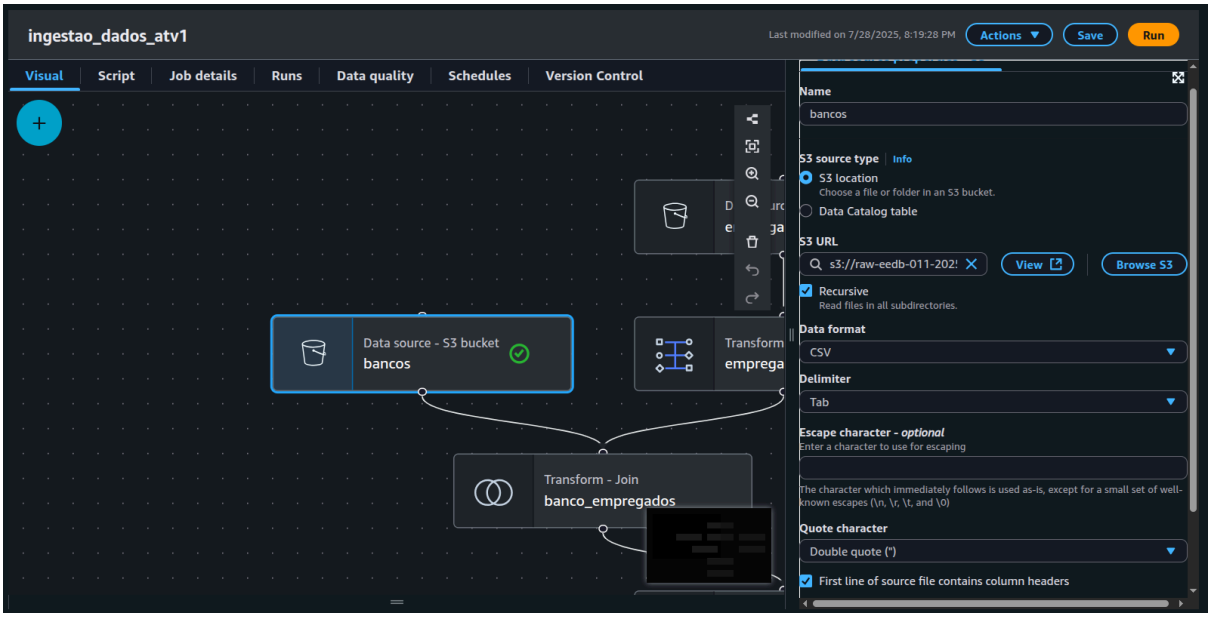
Passo 4 - Tratamento e junção dos dados utilizando ETL Visual (Glue)

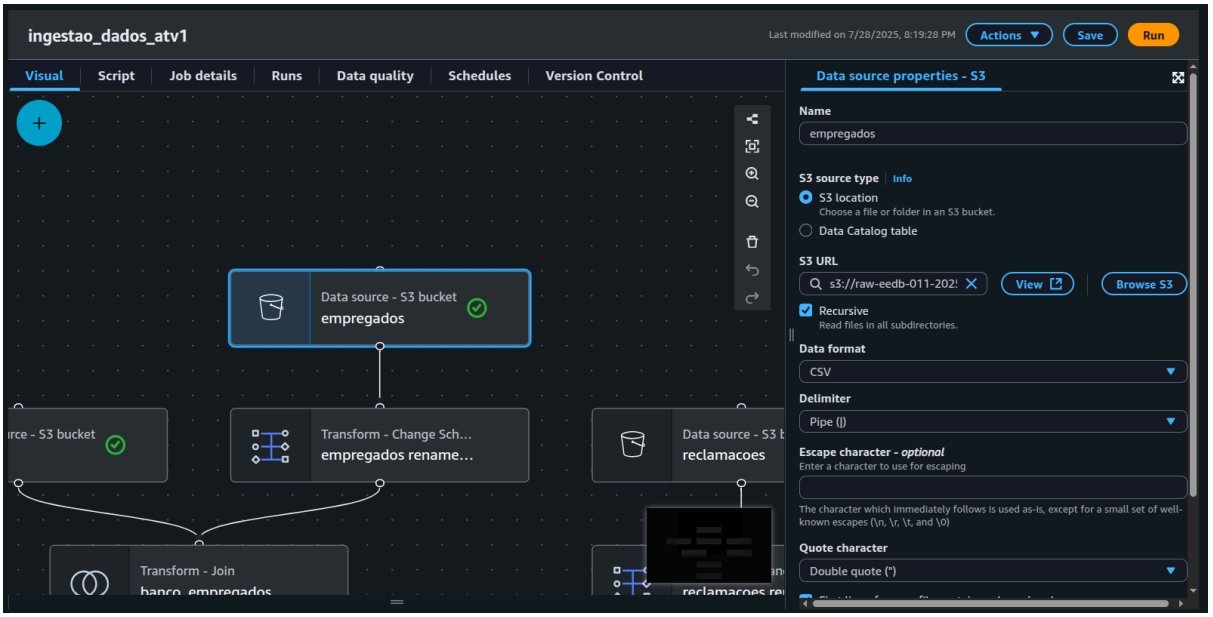
(a) Lendo as três tabelas diretamente do S3

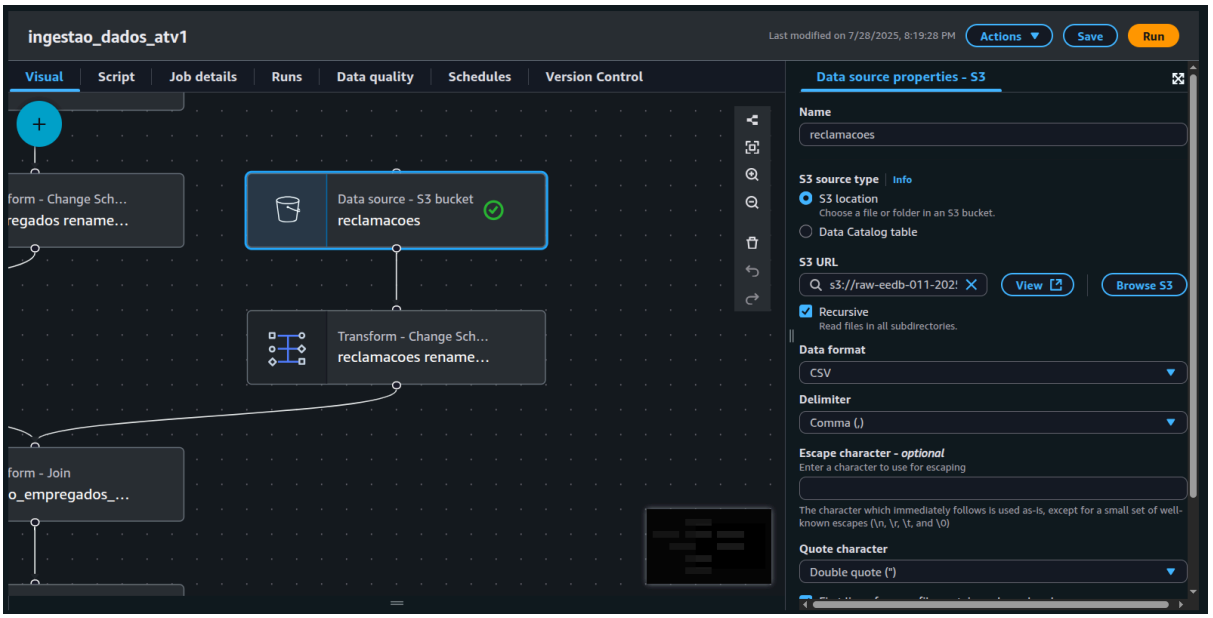
Inicialmente, realizamos a leitura dos dados das três tabelas diretamente do S3. Posteriormente, foram realizadas algumas operações de limpeza e transformações dos dados. E finalmente a união dos dados.



Parâmetros utilizados na leitura.







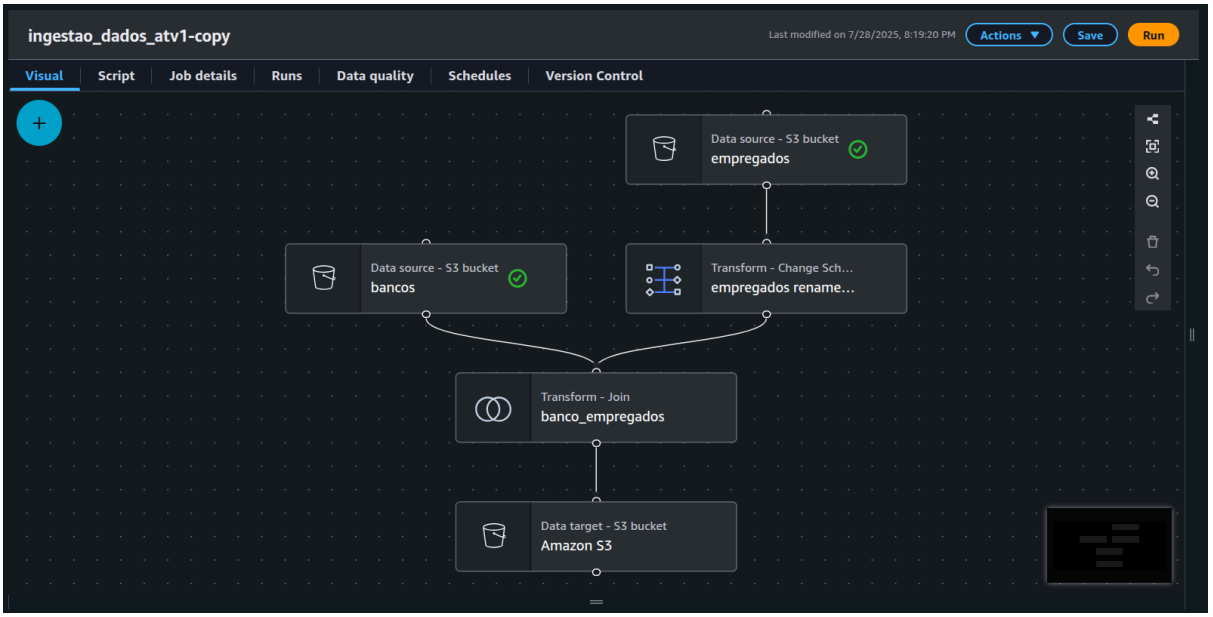
Resultado

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

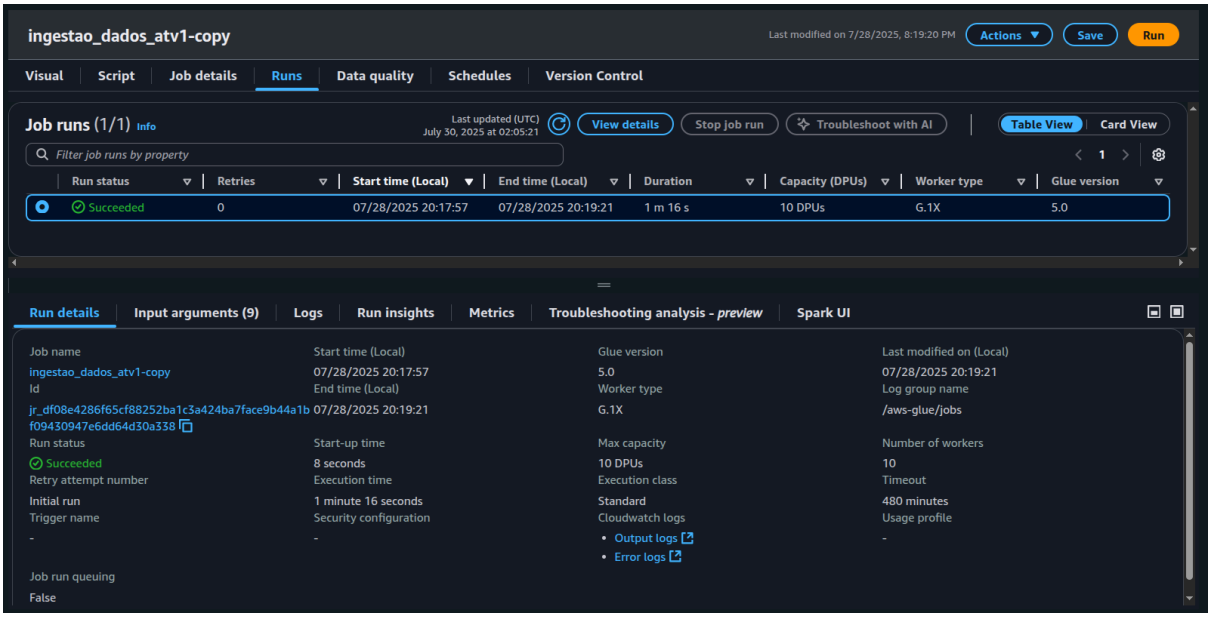
(b) Lendo as duas tabelas diretamente do S3

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

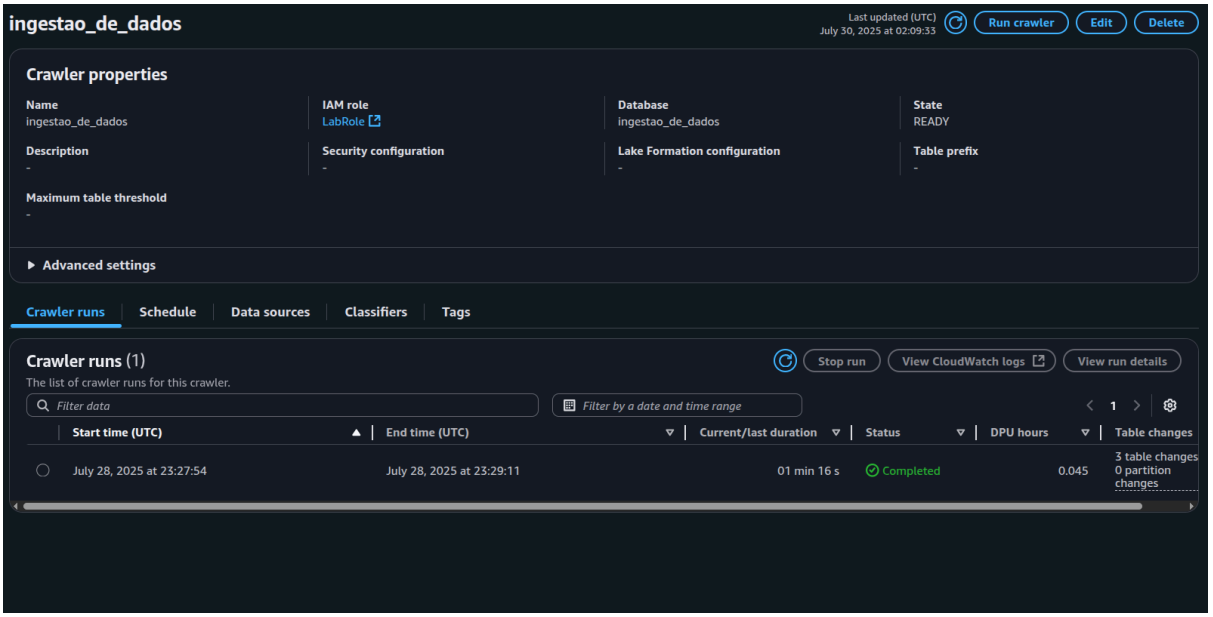
Como é possível observar na imagem a seguir, o job funcionou corretamente. Na seção seguinte será possível observar os dados resultantes.



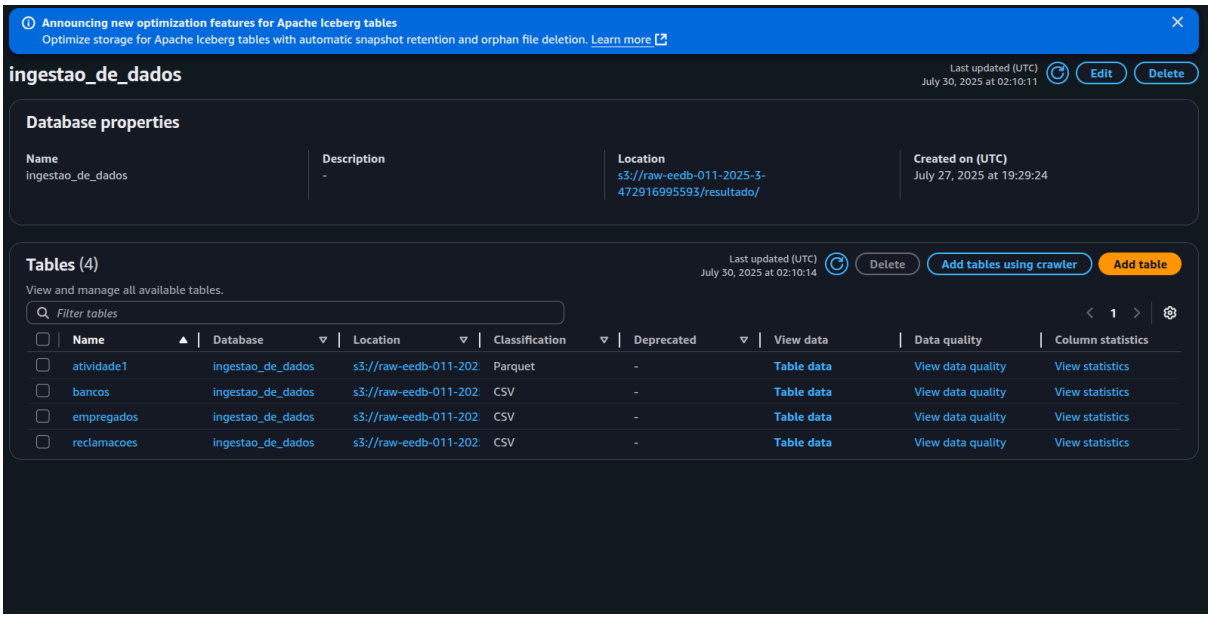
(c) Lendo as três tabelas diretamente do Catálogo

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.

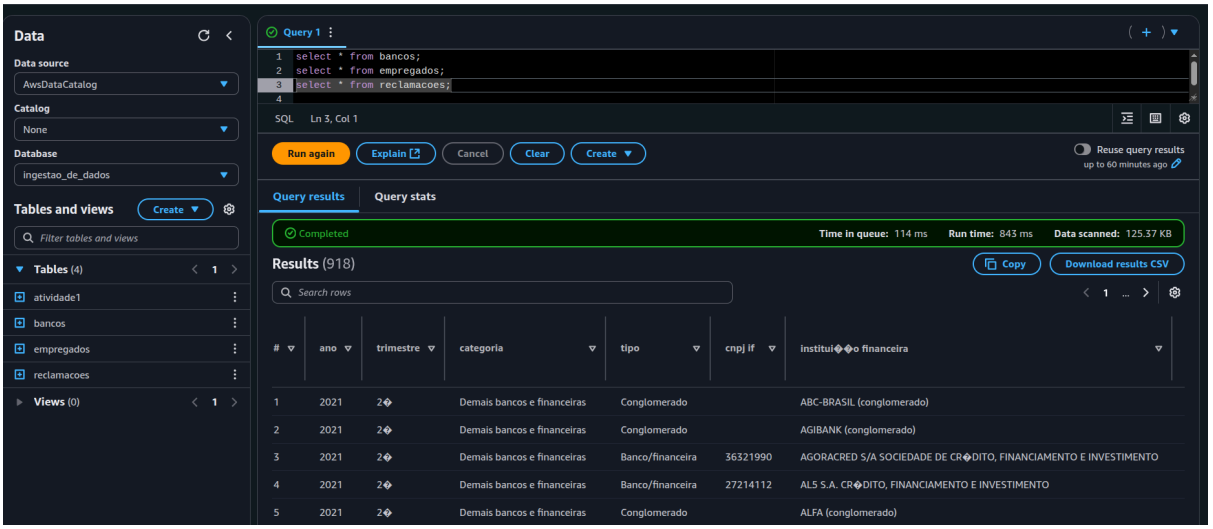
Crawler



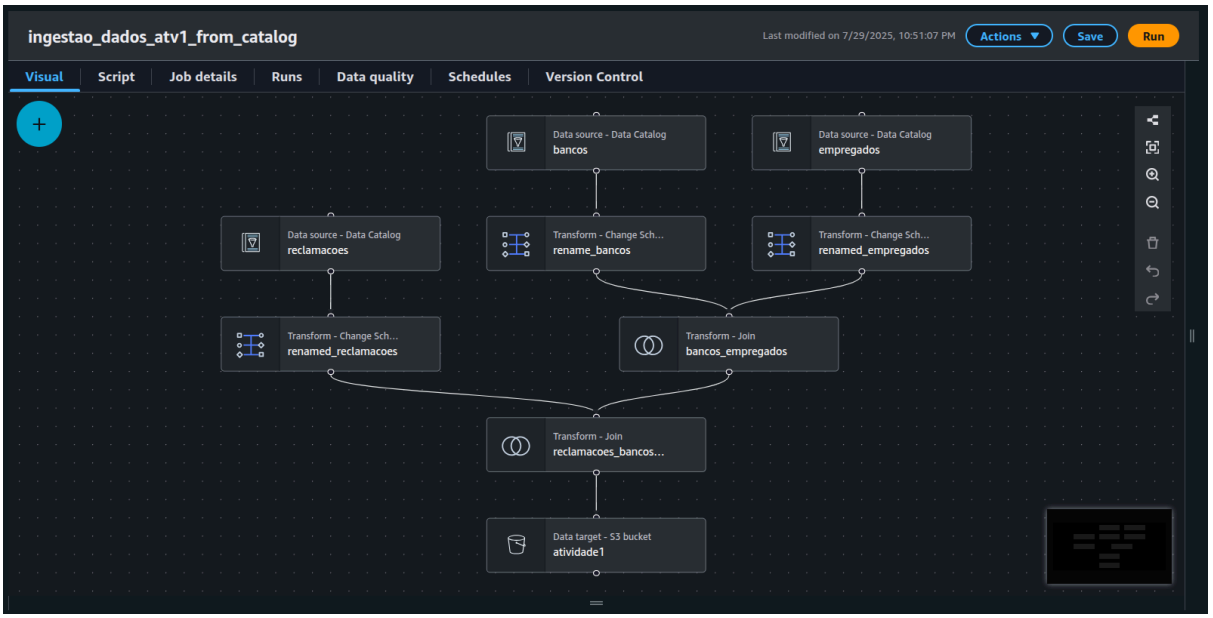
Catálogo

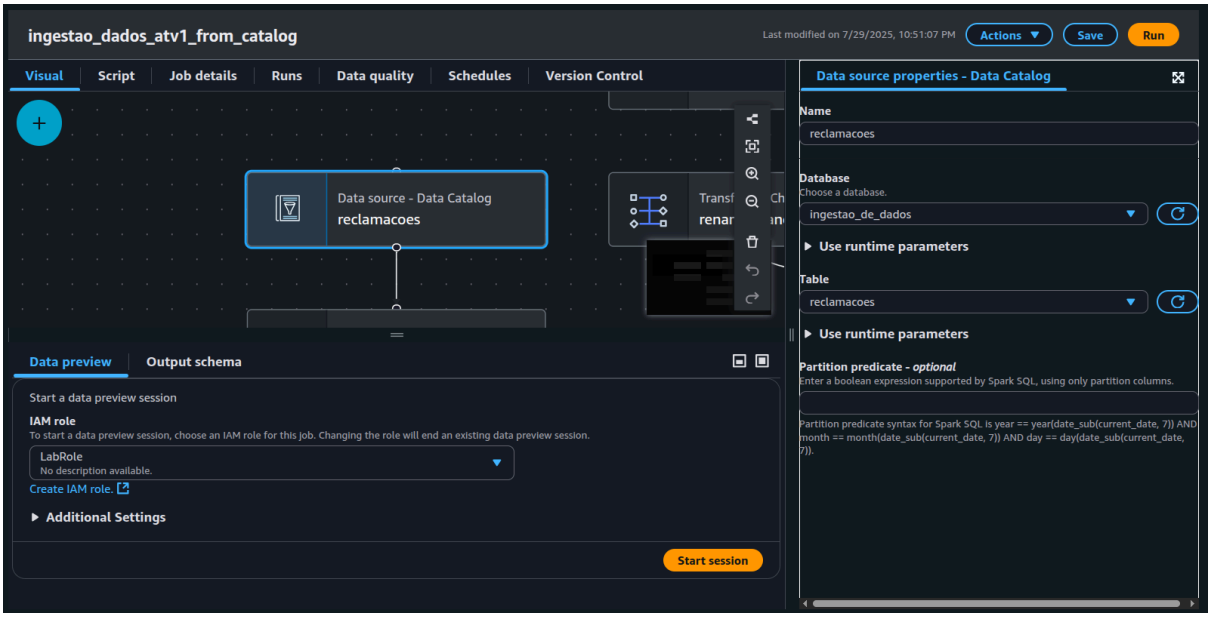


Athena - É ler os dados de reclamação catalogados através do Athena.


Data
Data source
AwsDataCatalog
Catalog
None
Database
ingestao_de_dados
Tables and views
Create
Filter tables and views
Tables (4)
atividade1
bancos
empregados
reclamacoes
Views (0)
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 cnpj if institui��o financeira
1 2021 2T Demais bancos e financeiras Conglomerado ABC-BRASIL (conglomerado)
2 2021 2T Demais bancos e financeiras Conglomerado AGIBANK (conglomerado)
3 2021 2T Demais bancos e financeiras Banco/financeira 36321990 AGORACRED S/A SOCIEDADE DE CR�DITO, FINANCIAMENTO E INVESTIMENTO
4 2021 2T Demais bancos e financeiras Banco/financeira 27214112 AL5 S.A. CR�DITO, FINANCIAMENTO E INVESTIMENTO
5 2021 2T Demais bancos e financeiras Conglomerado ALFA (conglomerado)

Job


ingestao_dados_atv1_from_catalog
Last modified on 7/29/2025, 10:51:07 PM
Actions
Save
Run
Visual
Script
Job details
Runs
Data quality
Schedules
Version Control
Data source - Data Catalog
bancos
Data source - Data Catalog
empregados
Data source - Data Catalog
reclamacoes
Transform - Change Sch...
rename_bancos
Transform - Change Sch...
renamed_empregados
Transform - Change Sch...
renamed_reclamacoes
Transform - Join
bancos_empregados
Transform - Join
reclamacoes_bancos...
Data target - S3 bucket
atividade1



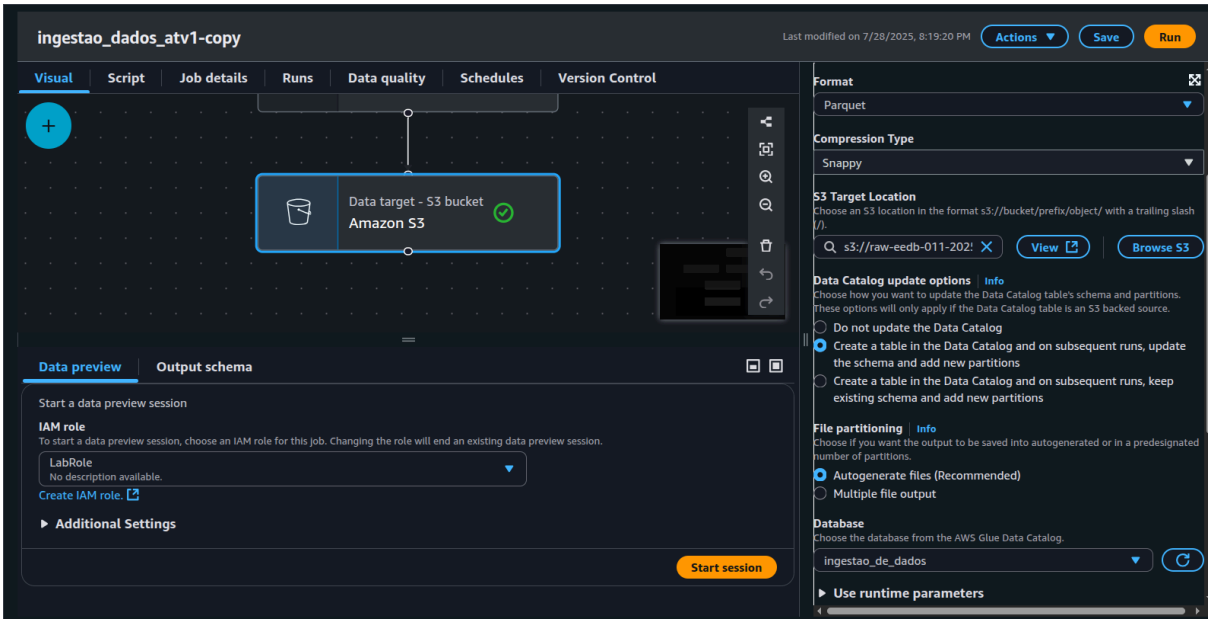
Resultado

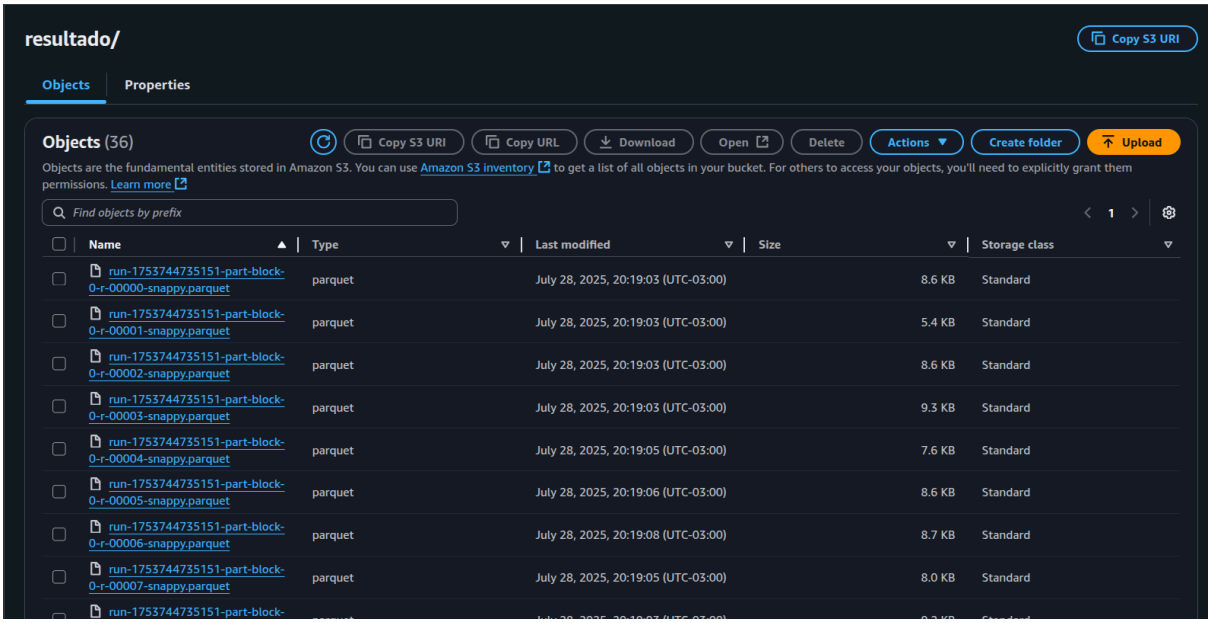
Ao tentar rodar o job ocorreu o mesmo erro para que impossibilitou a leitura dos dados de reclamação:

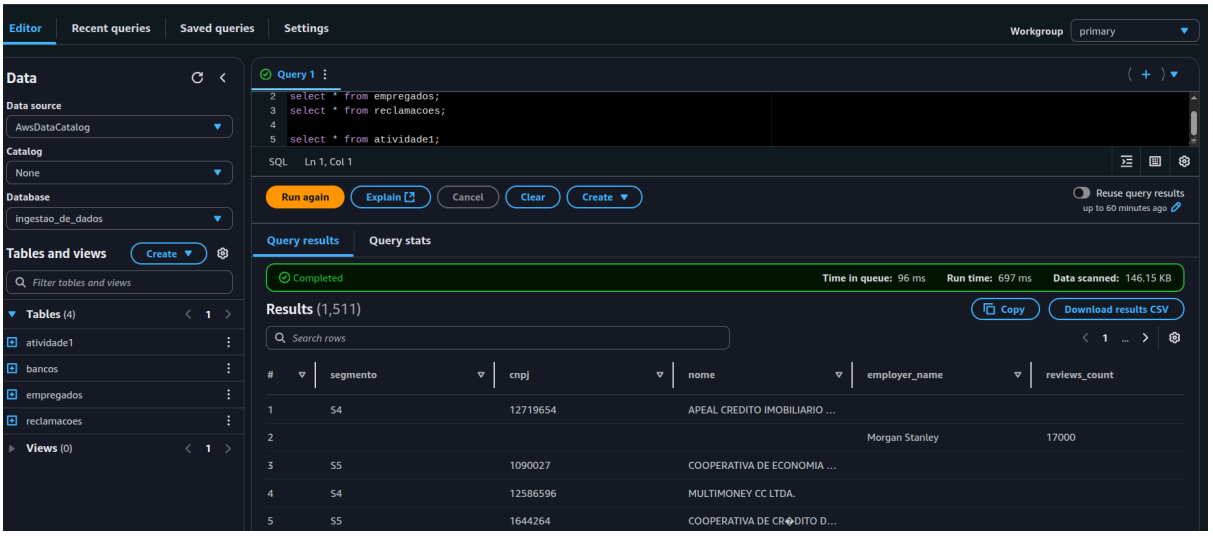
- Unable to parse file: 2022_tri_01.csv

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 os acesso aos dados através do Athena.


ingestao_dados_atv1-copy
Last modified on 7/28/2025, 8:19:20 PM
Actions
Save
Run
Visual
Script
Job details
Runs
Data quality
Schedules
Version Control
Data target - S3 bucket
Amazon S3
Format
Parquet
Compression Type
Snappy
S3 Target Location
Choose an S3 location in the format s3://bucket/prefix/object/ with a trailing slash (/).
s3://raw-eedb-011-202!
View
Browse S3
Data Catalog update options
Info
Choose how you want to update the Data Catalog table's schema and partitions. These options will only apply if the Data Catalog table is an S3 backed source.
Do not update the Data Catalog
Create a table in the Data Catalog and on subsequent runs, update the schema and add new partitions
Create a table in the Data Catalog and on subsequent runs, keep existing schema and add new partitions
File partitioning
Info
Choose if you want the output to be saved into autogenerated or in a predesignated number of partitions.
Autogenerate files (Recommended)
Multiple file output
Database
Choose the database from the AWS Glue Data Catalog.
ingestao_de_dados
Use runtime parameters
Data preview
Output schema
Start a data preview session
IAM role
To start a data preview session, choose an IAM role for this job. Changing the role will end an existing data preview session.
LabRole
No description available.
Create IAM role.
Additional Settings
Start session

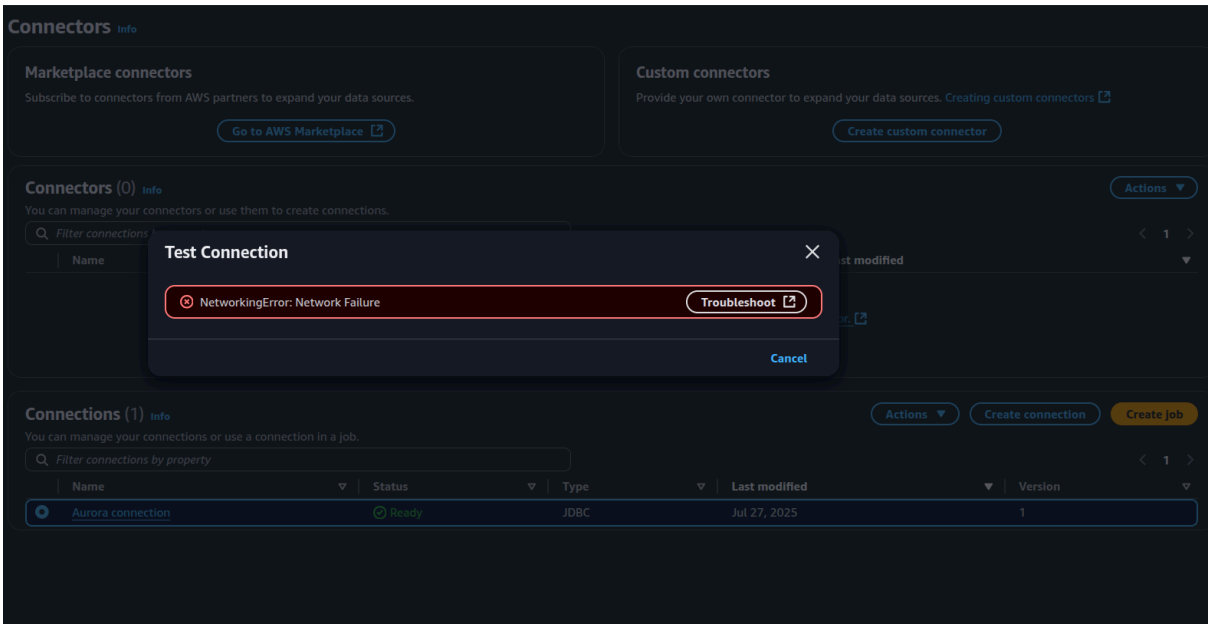

resultado/
Copy S3 URI
Objects
Properties
Objects (36)
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more
Find objects by prefix
Name
Type
Last modified
Size
Storage class
run-1753744735151-part-block-0-r-00000-snappy.parquet parquet July 28, 2025, 20:19:03 (UTC-03:00) 8.6 KB Standard
run-1753744735151-part-block-0-r-00001-snappy.parquet parquet July 28, 2025, 20:19:03 (UTC-03:00) 5.4 KB Standard
run-1753744735151-part-block-0-r-00002-snappy.parquet parquet July 28, 2025, 20:19:03 (UTC-03:00) 8.6 KB Standard
run-1753744735151-part-block-0-r-00003-snappy.parquet parquet July 28, 2025, 20:19:03 (UTC-03:00) 9.3 KB Standard
run-1753744735151-part-block-0-r-00004-snappy.parquet parquet July 28, 2025, 20:19:05 (UTC-03:00) 7.6 KB Standard
run-1753744735151-part-block-0-r-00005-snappy.parquet parquet July 28, 2025, 20:19:06 (UTC-03:00) 8.6 KB Standard
run-1753744735151-part-block-0-r-00006-snappy.parquet parquet July 28, 2025, 20:19:08 (UTC-03:00) 8.7 KB Standard
run-1753744735151-part-block-0-r-00007-snappy.parquet parquet July 28, 2025, 20:19:05 (UTC-03:00) 8.0 KB Standard


Editor
Recent queries
Saved queries
Settings
Workgroup
primary
Data
Data source
AwsDataCatalog
Catalog
None
Database
ingestao_de_dados
Tables and views
Create
Filter tables and views
Tables (4)
atividade1
bancos
empregados
reclamacoes
Views (0)
Query 1
select * from empregados;
select * from reclamacoes;
select * from atividade1;
SQL
Ln 1, Col 1
Run again
Explain
Cancel
Clear
Create
Reuse query results
up to 60 minutes ago
Query results
Query stats
Completed
Time in queue: 96 ms
Run time: 697 ms
Data scanned: 146.15 KB
Results (1,511)
Copy
Download results CSV
Search rows
segmento cnpj nome employer_name reviews_count
1 S4 12719654 APEAL CREDITO IMOBILIARIO ...
2 Morgan Stanley 17000
3 S5 1090027 COOPERATIVA DE ECONOMIA ...
4 S4 12586596 MULTIMONEY CC LTDA.
5 S5 1644264 COOPERATIVA DE CR?DITO D...

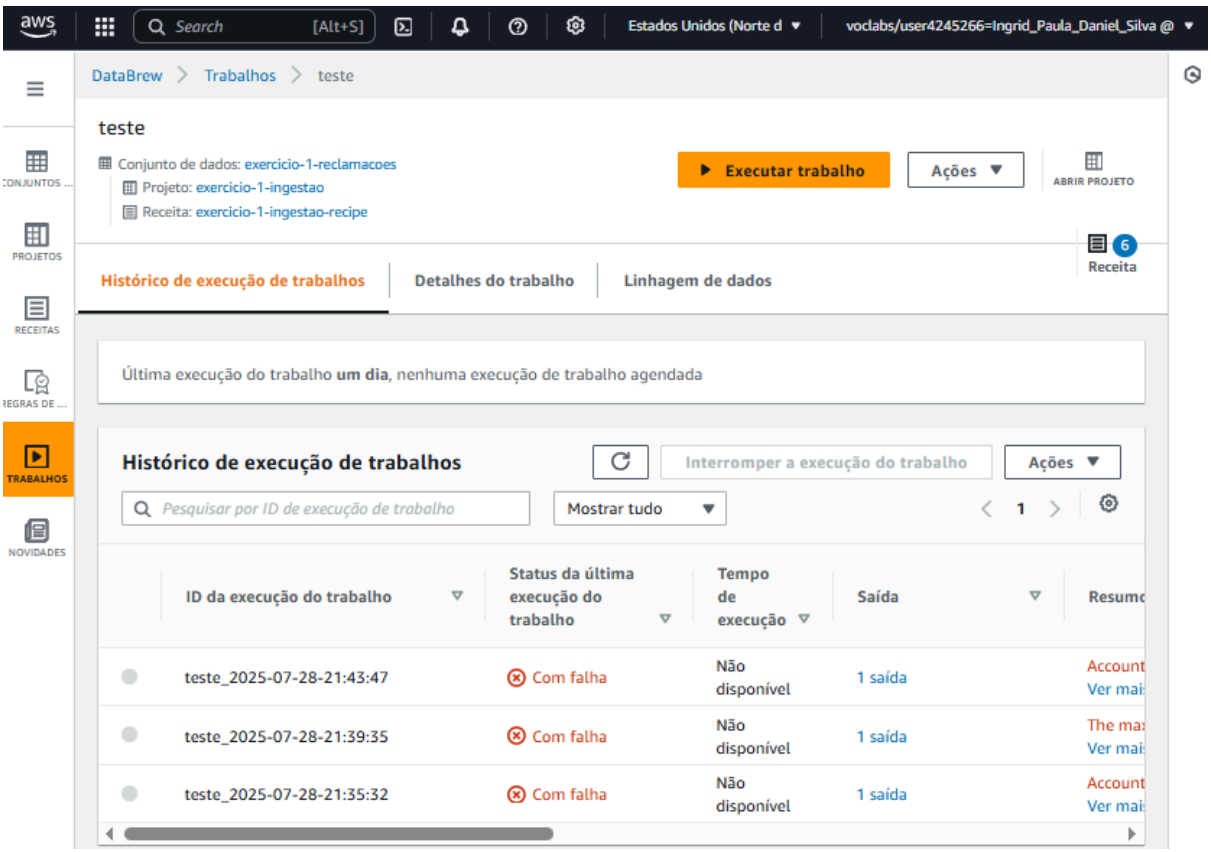
Extra - Bloqueios e desafios encontrados

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



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