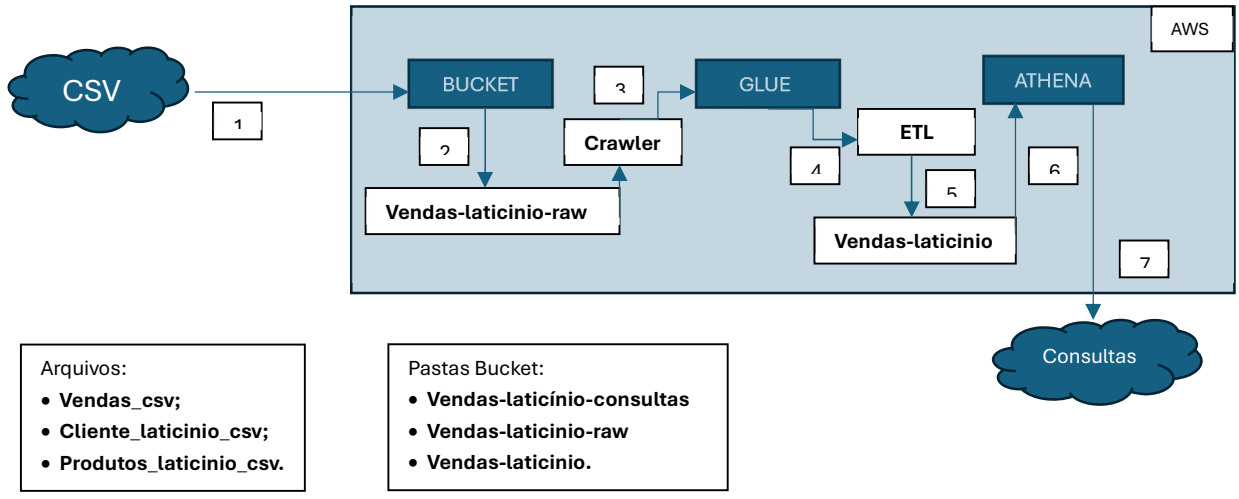
MVP Engenharia de Dados

1. Objetivo

A partir de dados de vendas entre 2023 e 2024 de um laticínio produtor de derivados de leite de cabra, analisar:

1. Quais produtos vendem mais em quantidade ?
2. Quais produtos vendem mais em peso ?
3. Quais produtos vendem mais em valor ?
4. Quais os principais clientes em valor de compras ?
5. Em quais Estados ocorrem as maiores vendas em valor ?
6. Em que meses do ano ocorrem as maiores vendas em valor ?

2. Arquitetura da solução



Arquivos:
- **Vendas_csv;**
- **Cliente_laticinio_csv;**
- **Produtos_laticinio_csv.**

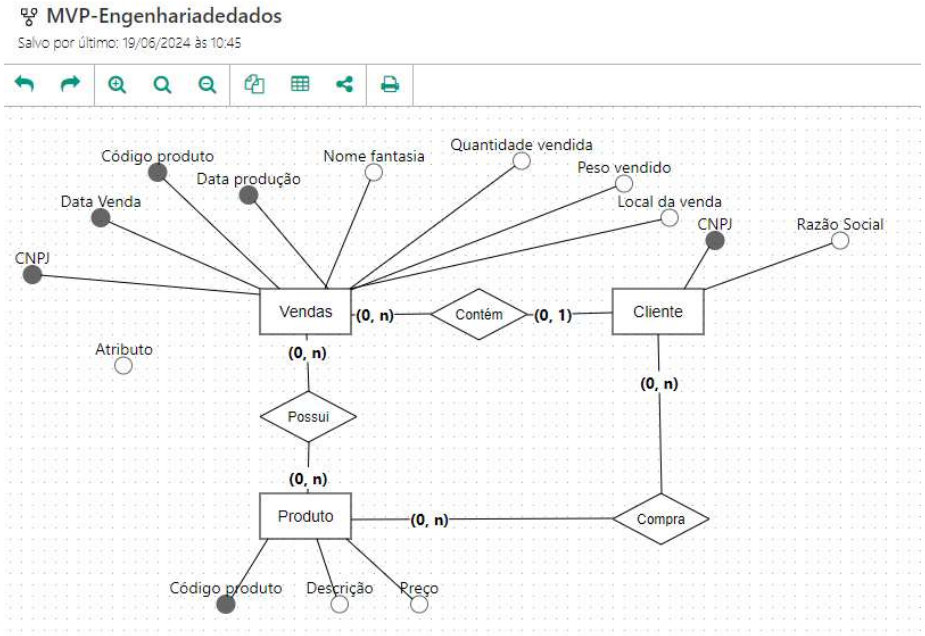
Pastas Bucket:
- **Vendas-laticínio-consultas**
- **Vendas-laticinio-raw**
- **Vendas-laticinio.**

1) Extraído do sistema de vendas do laticínio os arquivos em formato CSV de vendas, clientes e produtos;
2) Criação do Bucket e pastas no S3. Carga dos arquivos CSV (dados brutos) para pasta vendas-laticinio-raw;
3) Criação e execução do crawler no Glue para criação do Catálogo de dados das tabelas;
4) Desenho dos processos de DTL no Glue para criação de tabela consolidada;
5) Disparo do job no Glue para criação de tabela-consolidada (formato parquet) para pasta vendas-laticinio no S3;
6) Criação de queries SQL no AWS Athena para respostas aos objetivos do MVP;
7) Resultado das queries e análise.

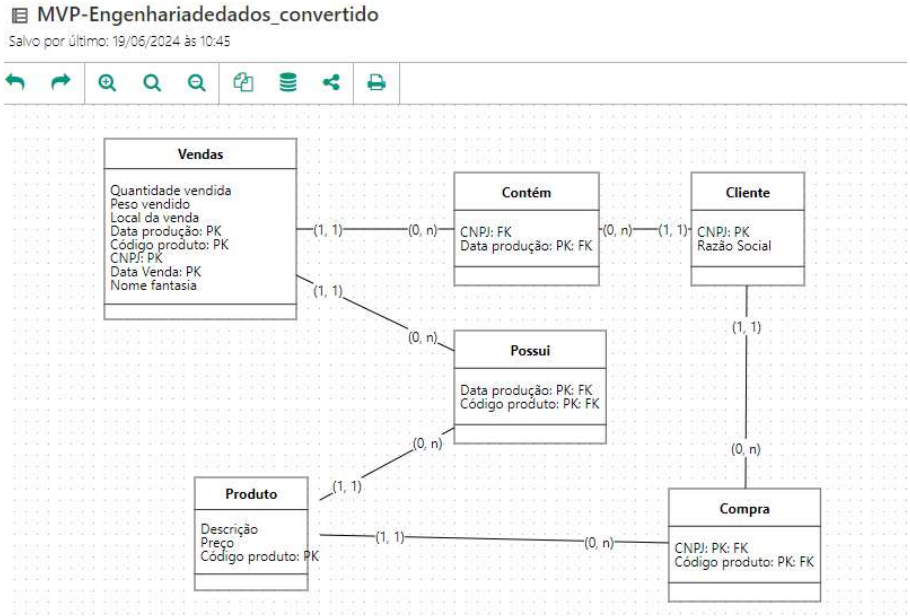
3. Modelagem

Para resposta ao problema foi modelada uma solução simples com 3 tabelas usando um esquema **Estrela**. As tabelas e atributos estão referenciadas no modelo conceitual e lógico abaixo criado pelo BR Modelo Web.

Modelo Conceitual



Modelo lógico



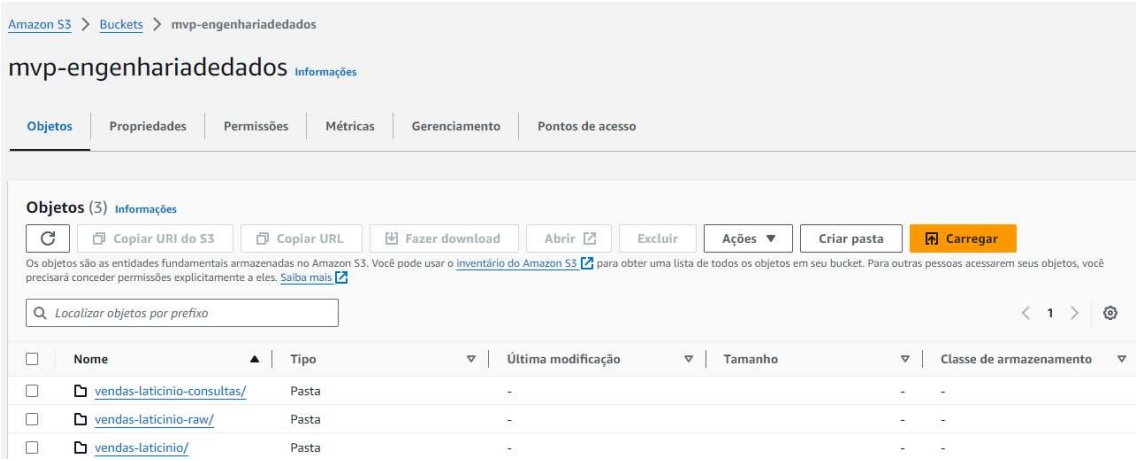
4. Detalhamento

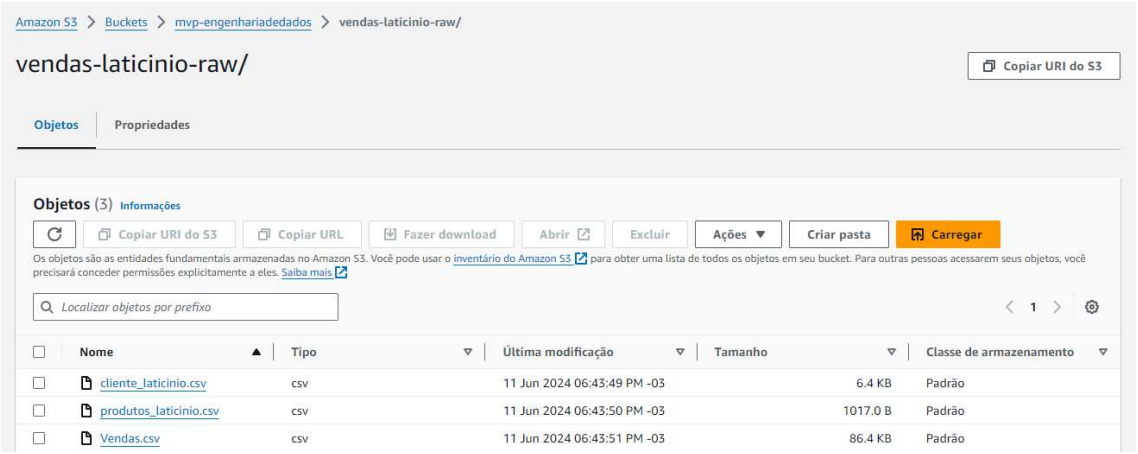
Extração em arquivos CSV de uma solução de controle de laticínios (arquivos na pasta do GitHub).

Obs: Concedo autorização para compartilhamento das informações por ser sócio-proprietário do empreendimento.

5. Coleta

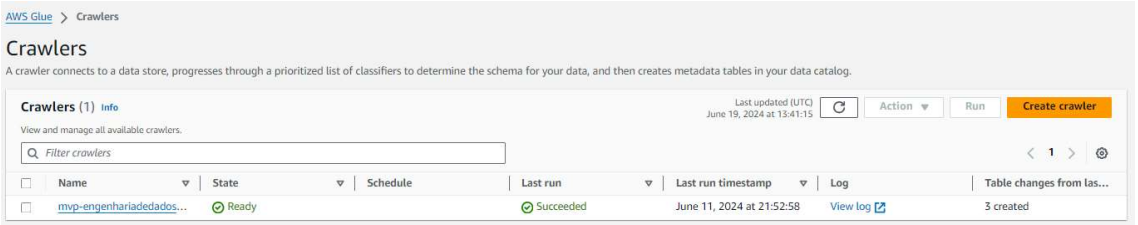
Dados baixados para máquina local e carga manual em um bucket do S3.

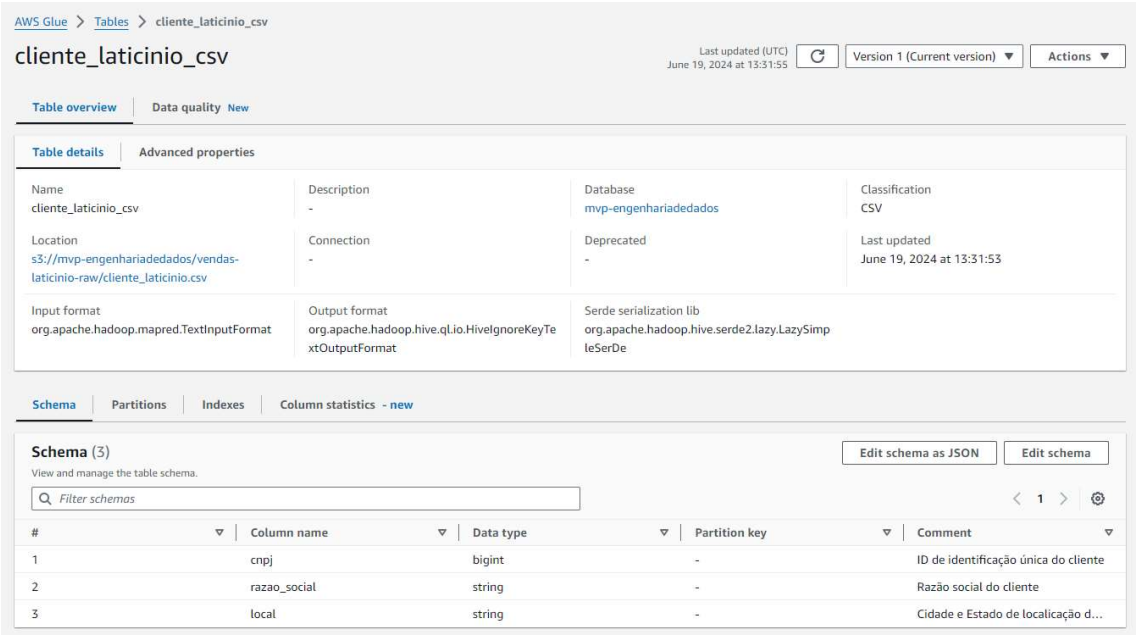


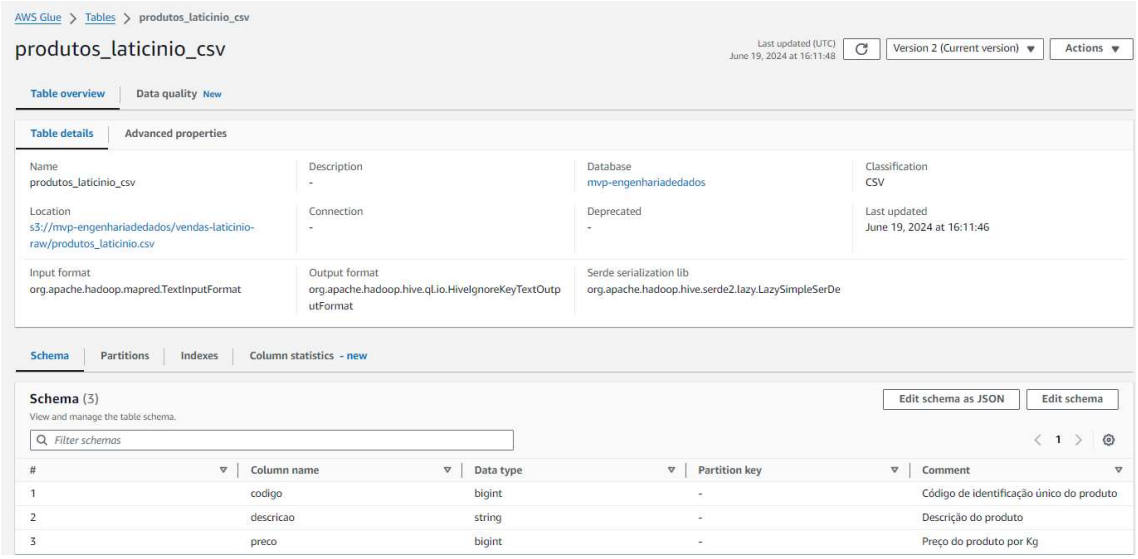


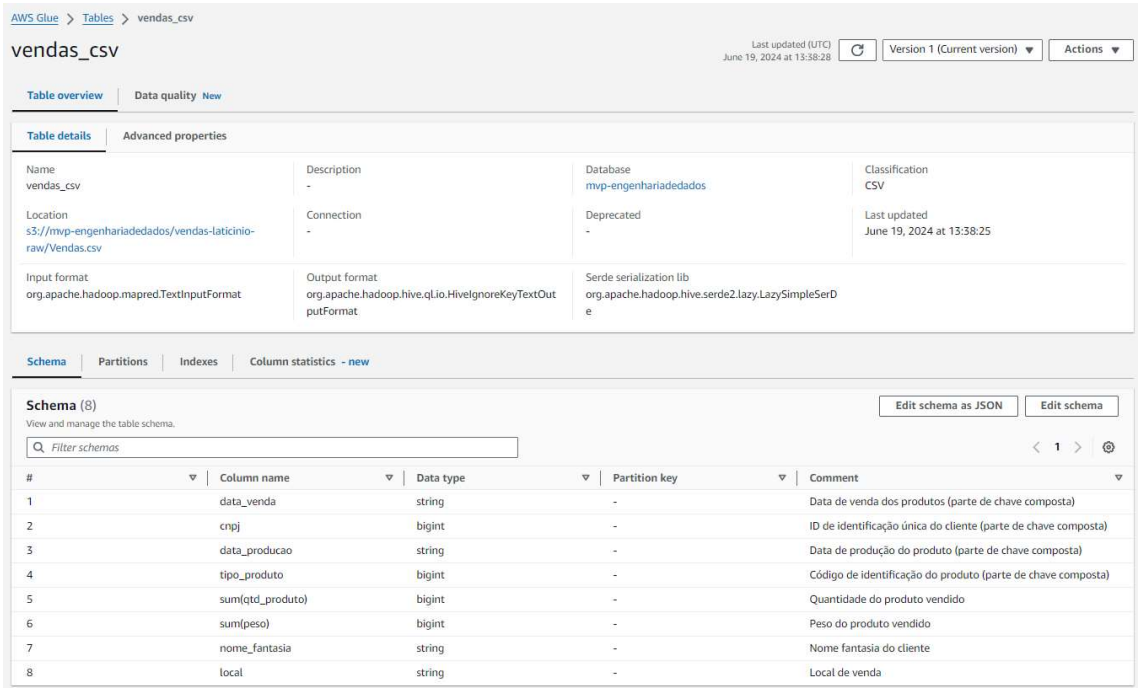
6. Criação do catálogo de dados

Utilizado o Crawler do AWS Glue para criação do Catálogo de dados









AWS Glue > Tables > vendas_csv

vendas_csv

Last updated (UTC)
June 19, 2024 at 13:38:28

Version 1 (Current version) | Actions

Table overview | Data quality New

Table details | Advanced properties

Name	Description	Database	Classification
vendas_csv	-	mvp-engenhariadedados	CSV
Location	**Connection**	**Deprecated**	**Last updated**
s3://mvp-engenhariadedados/vendas-laticinio-raw/Vendas.csv	-	-	June 19, 2024 at 13:38:25
Input format	**Output format**	**Serde serialization lib**	
org.apache.hadoop.mapred.TextInputFormat	org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat	org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe	

Schema | Partitions | Indexes | Column statistics - new

Schema (8)

View and manage the table schema.

Edit schema as JSON | Edit schema

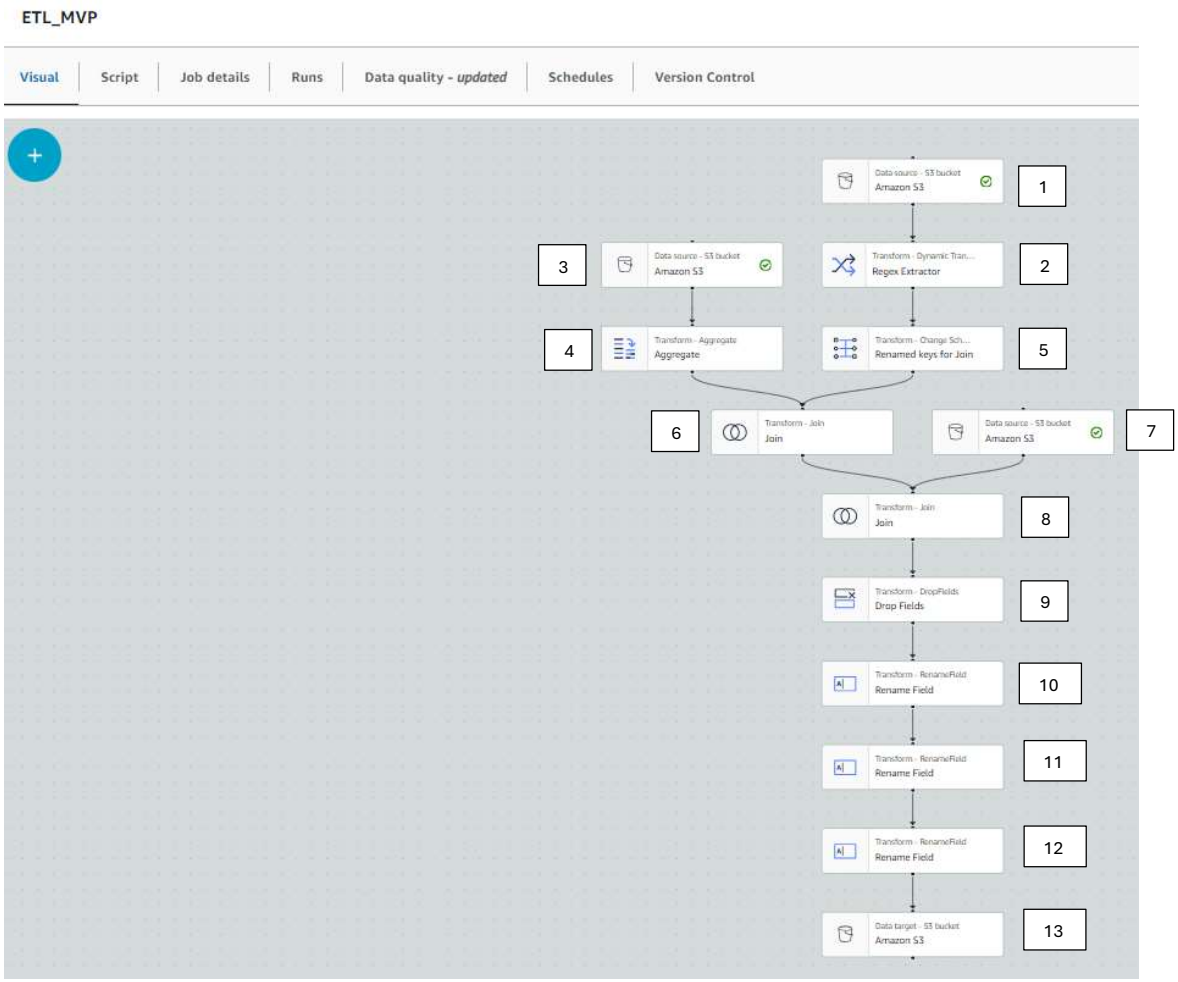
Filter schemas

1

#	Column name	Data type	Partition key	Comment
1	data_venda	string	-	Data de venda dos produtos (parte de chave composta)
2	cnpj	bigint	-	ID de identificação única do cliente (parte de chave composta)
3	data_producao	string	-	Data de produção do produto (parte de chave composta)
4	tipo_produto	bigint	-	Código de identificação do produto (parte de chave composta)
5	sum(qtd_produto)	bigint	-	Quantidade do produto vendido
6	sum(peso)	bigint	-	Peso do produto vendido
7	nome_fantasia	string	-	Nome fantasia do cliente
8	local	string	-	Local de venda

7. Carga e transformação dos dados

O ETL foi realizado utilizando o serviço AWS Glue. Foram criadas as seguintes etapas pela sua interface:

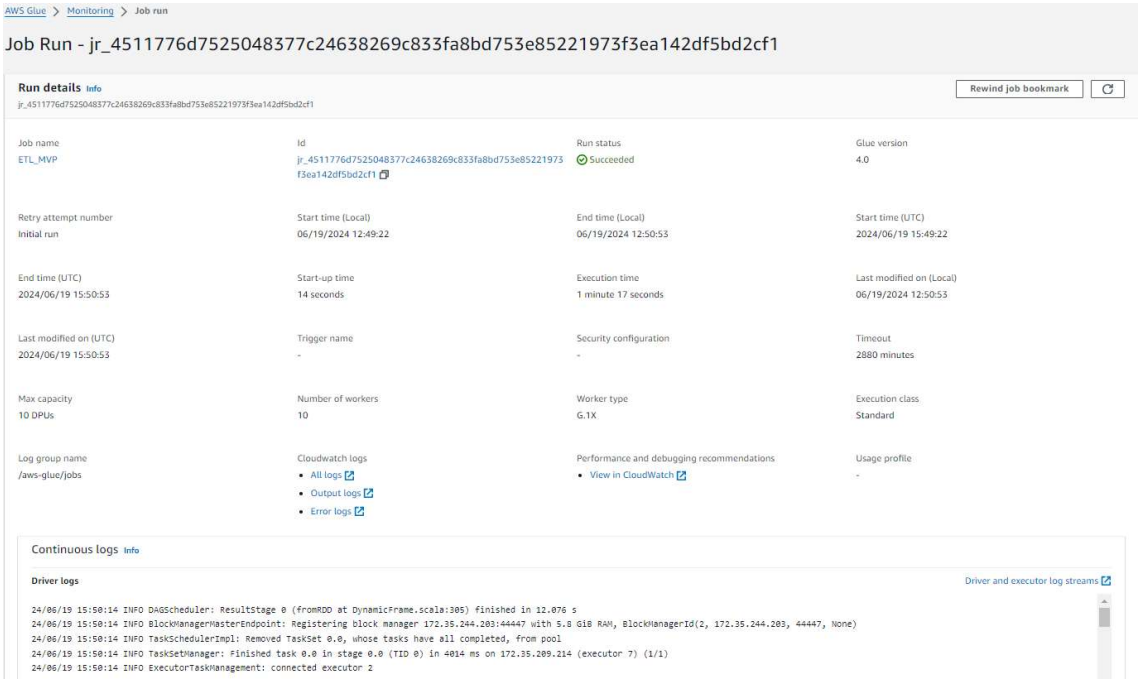


Etapas:

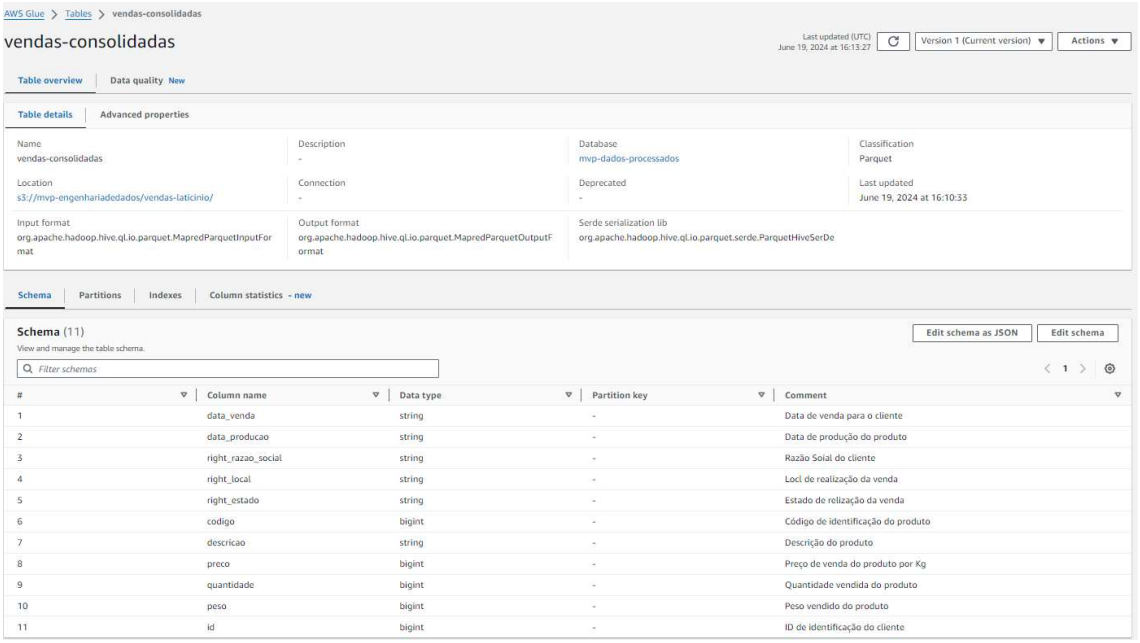
1) Extração dos dados de Clientes_laticino_csv utilizando o catálogo de dados criado pelo Crawler no BD mvp-engenhariadedados e existente no Bucket do S3;
2) Criação de uma coluna Estado a partir da coluna local;
3) Extração dos dados de vendas_csv utilizando o catálogo de dados criado pelo Crawler no BD mvp-engenhariadedados e existente no Bucket do S3;
4) Realizar agregação dos dados de venda por Data_Venda, CNPJ, Data_Produção e Tipo_Produto (Consolidar todas as vendas de um determinado cliente por tipo de produto adquirido e sua data de produção);
5) Renomear colunas;
6) Join de tabelas vendas_csv agregada com os clientes;
7) Extração dos dados de produtos_laticinio_csv utilizando o catálogo de dados criado pelo Crawler no BD mvp-engenhariadedados e existente no Bucket do S3;
8) Join da tabela criada em 6 com a tabela de produtos_laticinio_csv;
9) Remoção de colunas redundantes;
10) Renomear coluna agregada (somatório da quantidade de produto) para quantidade;
11) Renomear coluna agregada (somatório do peso do produto) para peso;

12) Renomear coluna agregada (CNPJ) para ID;
13) Carga de dados na tabela no BD mvp-dados-processados e criação do catálogo de dados com o nome vendas-consolidadas em formato parquet e gravação no Bucket do S3 na pasta vendas_laticinio.

Resultado Job executado



Catálogo de dados para tabela consolidada criada



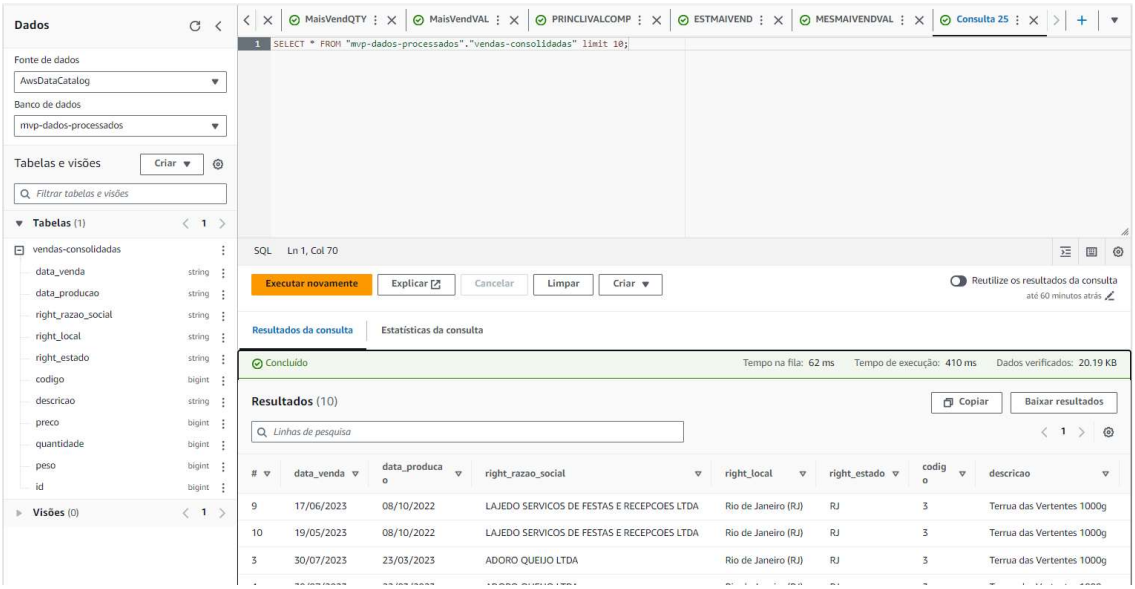
8. Análise

Foram criadas queries no AWS Athena utilizando a tabela consolidade pelo ETL para resposta as questões propostas. Os resultados das queries estão salvos no GitHub.

Os dados das tabelas foram tratados no sistema de origem, portanto não houve necessidade de nenhuma análise mais profunda para dados faltantes ou inconsistentes. Apenas houve a necessidade de separação de colunas como no caso da localidade para obter o Estado e eliminação de colunas redundantes.

A coluna CNPJ das tabelas pré-transformação não segue totalmente a lei de formação do Governo, é apenas um identificador único numérico que pode conter um CNPJ (validado no sistema de origem) ou um código. Explicado no catálogo de dados e renomeado após a execução do job ETL.

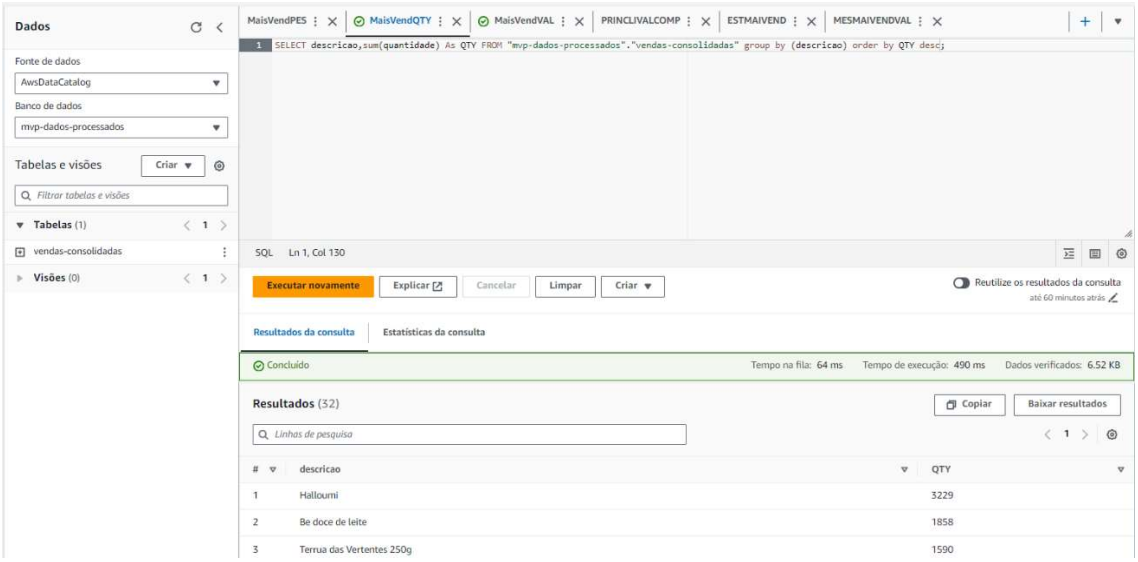
Tabela consolidada criada pelo ETL ao executar o Job no BD mvp-dados-processados:



Respostas:

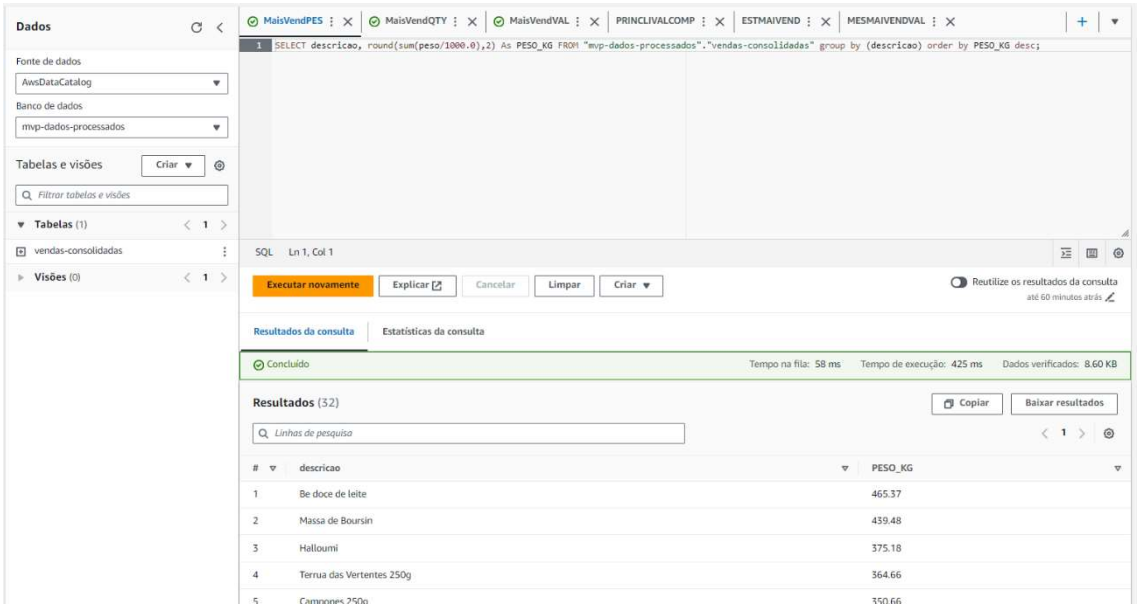
1. Quais produtos vendem mais em quantidade ?

   Criada Querie **MaisVendQTY**



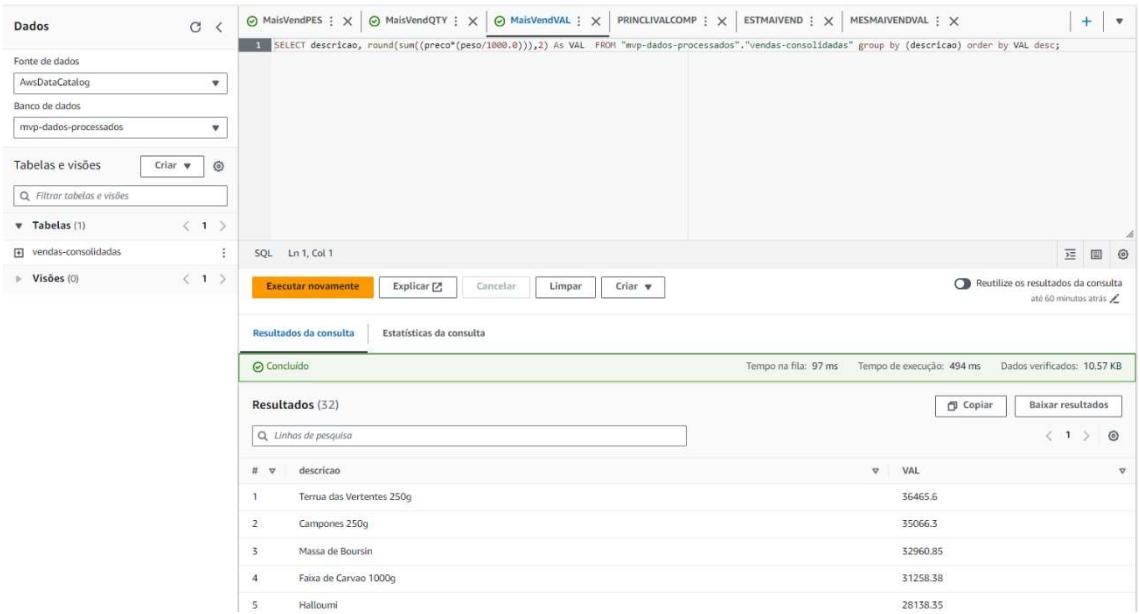
2. Quais produtos vendem mais em peso ?

Criada Querie **MaisVendPES**



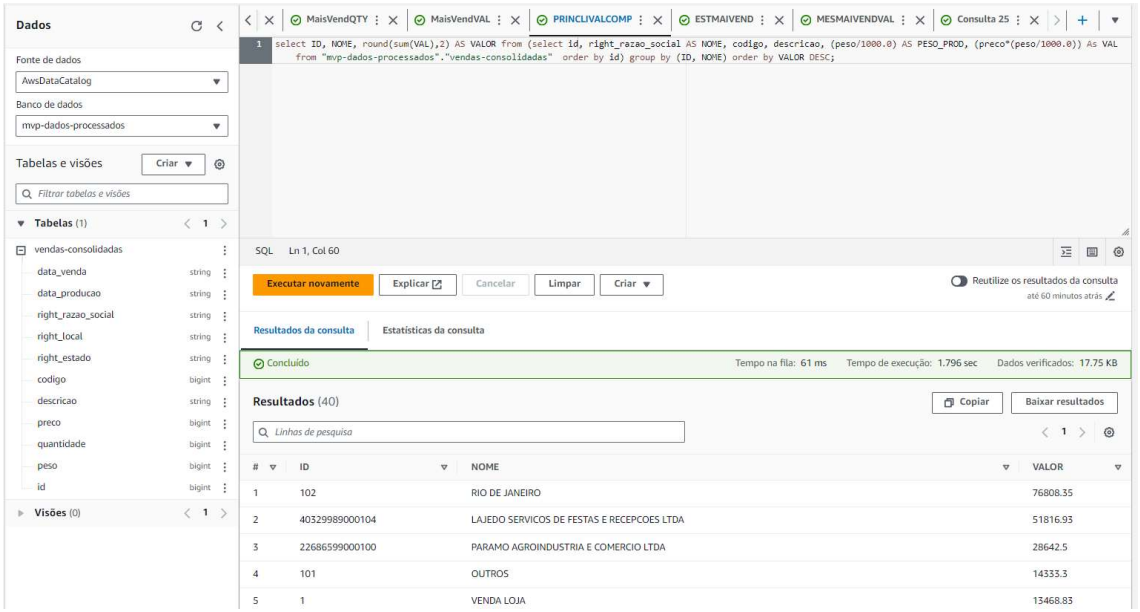
3. Quais produtos vendem mais em valor ?

Criada Querie **MaisVendVAL**



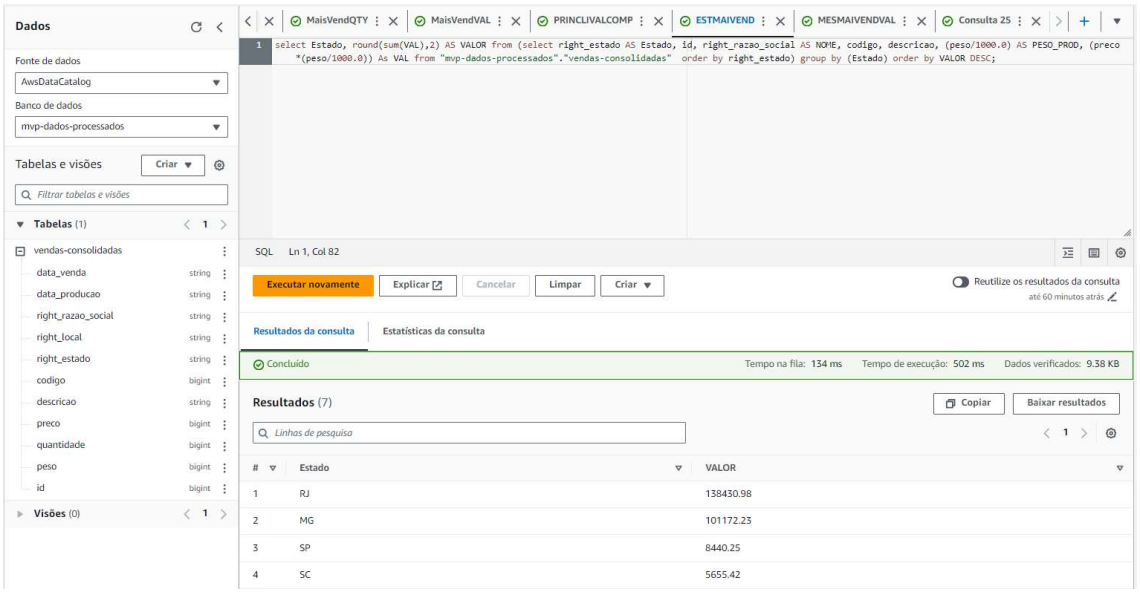
4. Quais os principais clientes em valor de compras ?

Criada Querie **PRINCLIVALCOMP**



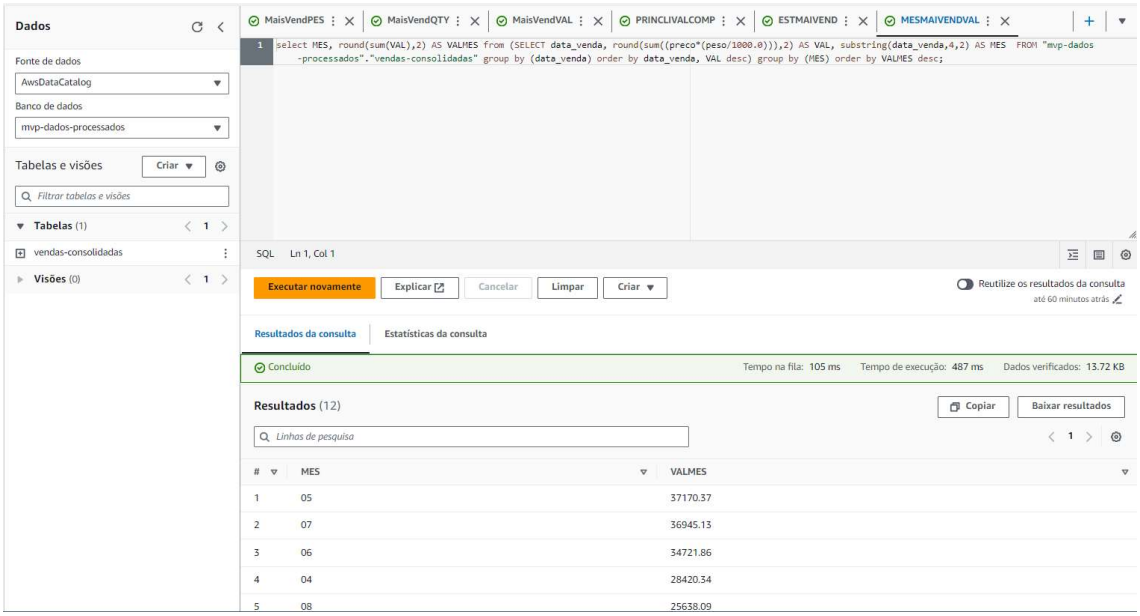
5. Em quais Estados ocorrem as maiores vendas em valor ?

Criada Querie **ESTMAIVEND**



6. Em que meses do ano ocorrem as maiores vendas em valor ?

Criada Querie **MESMAIVENDVAL**



O objetivo proposto foi baseado em descobertas para definição de estratégica comercial e de produção do laticínio. As respostas foram obtidas de forma simples em queries SQL após seguir todos os processos de modelagem, disponibilização dos arquivos CSV, carga, transformação e finalmente criar as consultas.

A plataforma AWS permitiu seguir todo o fluxo de forma simples e de fácil configuração com Buckets para os arquivos pelo S3, criação de Catálogo de dados e transformação pelo Glue e visualização/consultas pelo Athena.

O trabalho foi um protótipo para busca de respostas e futuramente criar um Data Warehouse ou Data Lake mais robusto, mas que serviu para um excelente entendimento de todo o processo.