

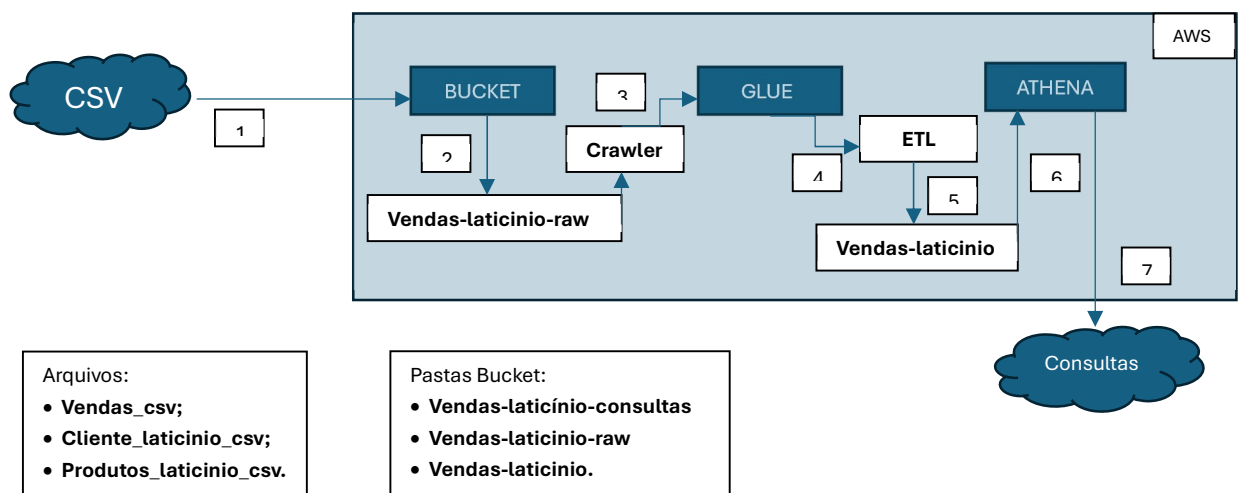
## MVP Engenharia de Dados

### 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

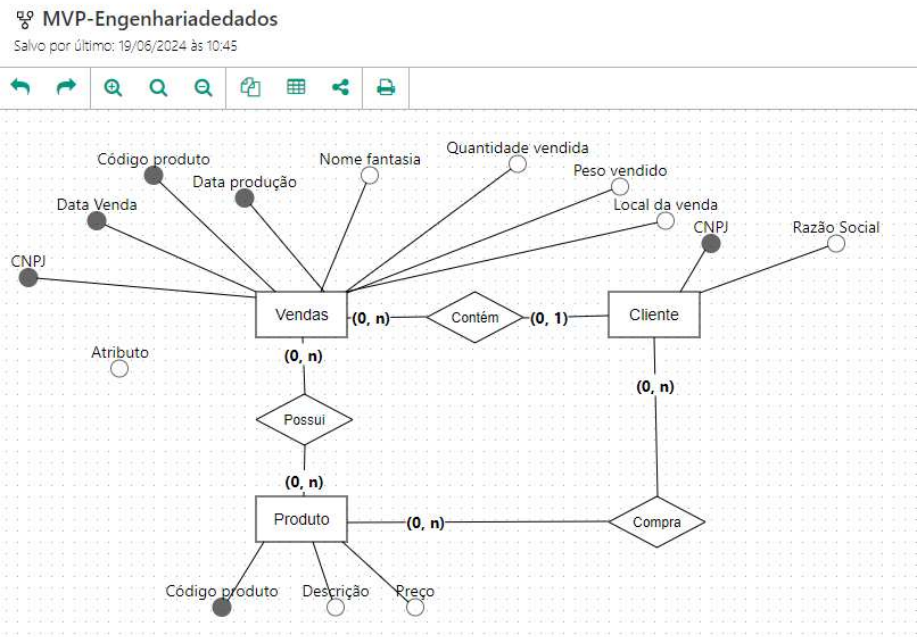


- 1) Extraído do sistema de vendas do laticínio os arquivos em formato CSV de vendas, clientes e produtos;
- 2) Criação do Bucket e pastas no S3. Carga dos arquivos CSV (dados brutos) para pasta vendas-laticinio-raw;
- 3) Criação e execução do crawler no Glue para criação do Catálogo de dados das tabelas;
- 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.

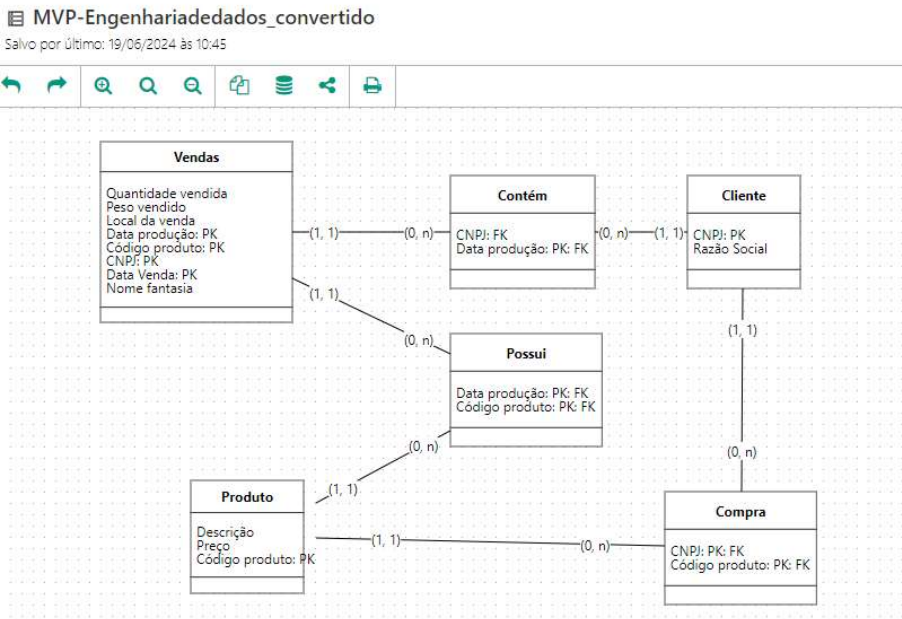
### 3. Modelagem

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.

Amazon S3 > Buckets > mvp-engenhariadedados

mvp-engenhariadedados

ObjetosPropriedadesPermissõesMétricasGerenciamentoPontos de acesso

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

< 1 >

	Nome	Tipo	Última modificação	Tamanho	Classe de armazenamento
<input type="checkbox"/>	vendas-laticinio-consultas/	Pasta	-	-	-
<input type="checkbox"/>	vendas-laticinio-raw/	Pasta	-	-	-
<input type="checkbox"/>	vendas-laticinio/	Pasta	-	-	-

Amazon S3 > Buckets > mvp-engenhariadedados > vendas-laticinio-raw/

vendas-laticinio-raw/

Copiar URI do S3

ObjetosPropriedades

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

< 1 >

	Nome	Tipo	Última modificação	Tamanho	Classe de armazenamento
<input type="checkbox"/>	cliente_laticinio.csv	csv	11 Jun 2024 06:43:49 PM -03	6.4 KB	Padrão
<input type="checkbox"/>	produtos_laticinio.csv	csv	11 Jun 2024 06:43:50 PM -03	1017.0 B	Padrão
<input type="checkbox"/>	Vendas.csv	csv	11 Jun 2024 06:43:51 PM -03	86.4 KB	Padrão

## 6. Criação do catálogo de dados

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

[AWS Glue](#) > Crawlers

### Crawlers

A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

Crawlers (1) Info

Last updated (UTC)  
June 19, 2024 at 13:41:15

Refresh

Action

Run

Create crawler

View and manage all available crawlers.

Q Filter crawlers

< 1 > ⚙

<input type="checkbox"/>	Name	State	Schedule	Last run	Last run timestamp	Log	Table changes from las...
<input type="checkbox"/>	mvp-engenhariadedados...	Ready		Succeeded	June 11, 2024 at 21:52:58	<a href="#">View log</a>	3 created

[AWS Glue](#) > [Tables](#) > cliente\_laticinio\_csv

### cliente\_laticinio\_csv

Last updated (UTC)  
June 19, 2024 at 13:31:53

Refresh

Version 1 (Current version)

Actions

Table overview

Data quality New

Table details

Advanced properties

Name cliente_laticinio_csv	Description -	Database <a href="#">mvp-engenhariadedados</a>	Classification CSV
Location <a href="#">s3://mvp-engenhariadedados/vendas-laticinio-raw/cliente_laticinio.csv</a>	Connection -	Deprecated -	Last updated June 19, 2024 at 13:31:53
Input format org.apache.hadoop.mapred.TextInputFormat	Output format org.apache.hadoop.hive.qLio.HiveIgnoreKeyTextOutputFormat	Serde serialization lib org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe	

Schema

Partitions

Indexes

Column statistics - new

Schema (3)

View and manage the table schema.

Edit schema as JSON

Edit schema

Q Filter schemas

< 1 > ⚙

#	Column name	Data type	Partition key	Comment
1	cnpj	bigint	-	ID de identificação única do cliente
2	razao_social	string	-	Razão social do cliente
3	local	string	-	Cidade e Estado de localização d...

[AWS Glue](#) > [Tables](#) > produtos\_laticinio\_csv

### produtos\_laticinio\_csv

Last updated (UTC)  
June 19, 2024 at 16:11:48

Refresh

Version 2 (Current version)

Actions

Table overview

Data quality New

Table details

Advanced properties

Name produtos_laticinio_csv	Description -	Database <a href="#">mvp-engenhariadedados</a>	Classification CSV
Location <a href="#">s3://mvp-engenhariadedados/vendas-laticinio-raw/produtos_laticinio.csv</a>	Connection -	Deprecated -	Last updated June 19, 2024 at 16:11:46
Input format org.apache.hadoop.mapred.TextInputFormat	Output format org.apache.hadoop.hive.qLio.HiveIgnoreKeyTextOutputFormat	Serde serialization lib org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe	

Schema

Partitions

Indexes

Column statistics - new

Schema (3)

View and manage the table schema.

Edit schema as JSON

Edit schema

Q Filter schemas

< 1 > ⚙

#	Column name	Data type	Partition key	Comment
1	codigo	bigint	-	Código de identificação único do produto
2	descricao	string	-	Descrição do produto
3	preco	bigint	-	Preço do produto por Kg

Table overview

Data quality New

Table details

Advanced properties

Name

vendas\_csv

Location

s3://mvp-engenhariadedados/vendas-laticinio-raw/Vendas.csv

Input format

org.apache.hadoop.mapred.TextInputFormat

Description

-

Connection

-

Output format

org.apache.hadoop.hive ql.io.HiveIgnoreKeyTextOutputFormat

Database

mvp-engenhariadedados

Deprecated

-

Serde serialization lib

org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe

Classification

CSV

Last updated

June 19, 2024 at 13:38:25

Schema

Partitions

Indexes

Column statistics - new

Schema (8)

Edit schema as JSON

Edit schema

View and manage the table schema.

Q

Filter schemas

<

1

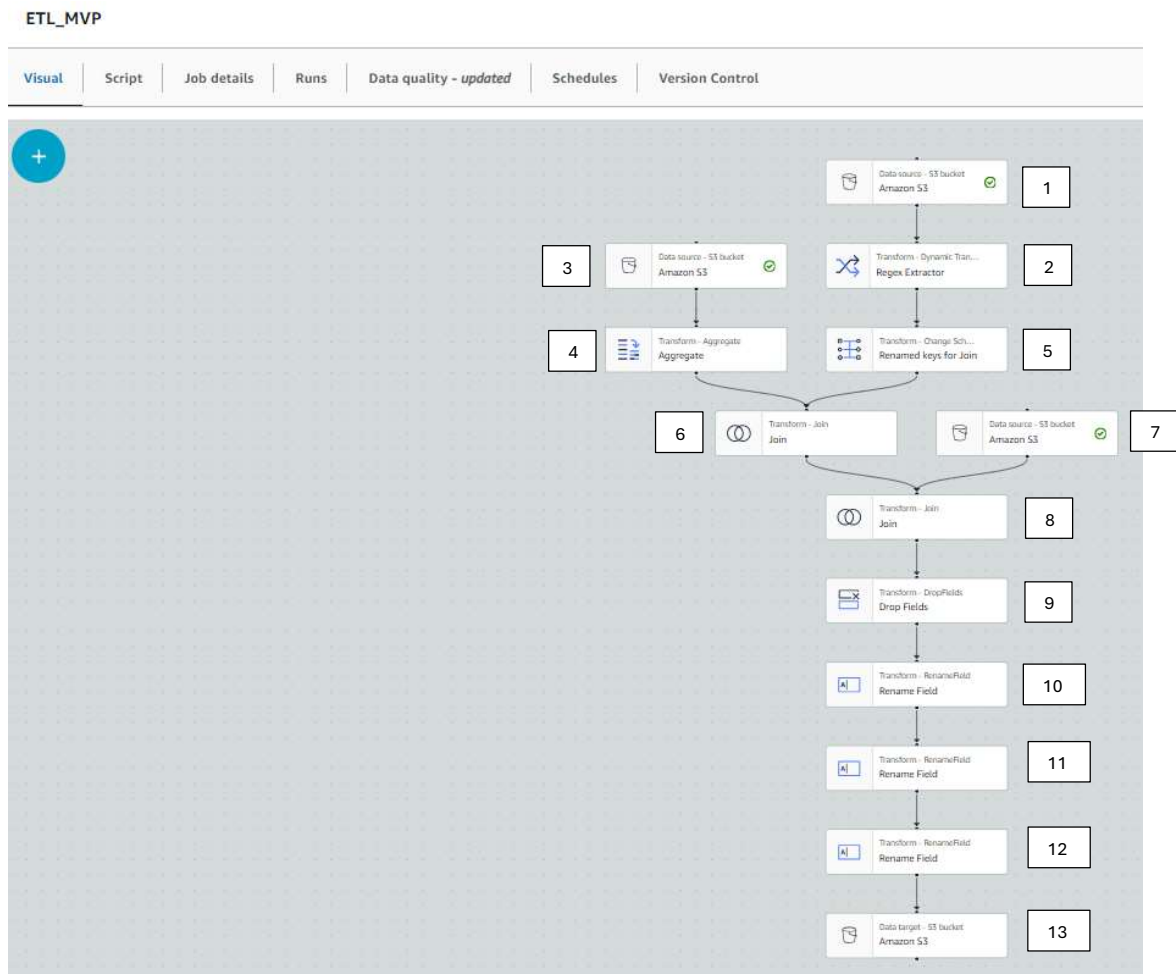
>

🔍

#	Column name	Data type	Partition key	Comment
1	data_venda	string	-	Data de venda dos produtos (parte de chave composta)
2	cnnpj	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

AWS Glue > Monitoring > Job run

Job Run - jr\_4511776d7525048377c24638269c833fa8bd753e85221973f3ea142df5bd2cf1

Run details info

Rewind job bookmark

jr\_4511776d7525048377c24638269c833fa8bd753e85221973f3ea142df5bd2cf1

Job name	Id	Run status	Glue version
ETL_MVP	jr_4511776d7525048377c24638269c833fa8bd753e85221973f3ea142df5bd2cf1	Succeeded	4.0
Retry attempt number	Start time (Local)	End time (Local)	Start time (UTC)
Initial run	06/19/2024 12:49:22	06/19/2024 12:50:53	2024/06/19 15:49:22
End time (UTC)	Start-up time	Execution time	Last modified on (Local)
2024/06/19 15:50:53	14 seconds	1 minute 17 seconds	06/19/2024 12:50:53
Last modified on (UTC)	Trigger name	Security configuration	Timeout
2024/06/19 15:50:53	-	-	2880 minutes
Max capacity	Number of workers	Worker type	Execution class
10 DPLUs	10	G.1X	Standard
Log group name	Cloudwatch logs	Performance and debugging recommendations	Usage profile
/aws-glue/jobs	<div>All logsOutput logsError logs</div>	<div>View in CloudWatch</div>	-

Continuous logs info

Driver logs

Driver and executor log streams

24/06/19 15:50:14 INFO DAGScheduler: ResultStage 0 (fromRDD at DynamicFrame.scala:305) Finished in 12.076 s  
24/06/19 15:50:14 INFO BlockManagerMasterEndpoint: Registering block manager 172.35.244.203:44447 with 5.8 GiB RAM, BlockManagerId(2, 172.35.244.203, 44447, None)  
24/06/19 15:50:14 INFO TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool  
24/06/19 15:50:14 INFO TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 4014 ms on 172.35.209.214 (executor 7) (1/1)  
24/06/19 15:50:14 INFO ExecutorTaskManagement: connected executor 2

Catálogo de dados para tabela consolidada criada

AWS Glue > Tables > vendas-consolidadas

ventas-consolidadas

Last updated (UTC)  
June 19, 2024 at 16:13:27

Version 1 (Current version)

Actions

Table overview

Data quality New

Table details

Advanced properties

Name	Description	Database	Classification
vendas-consolidadas	-	mvp-dados-processados	Parquet
Location	Connection	Deprecated	Last updated
s3://mvp-engenhariadados/vendas-laticinio/	-	-	June 19, 2024 at 16:10:33
Input format	Output format	Serde serialization lib	
org.apache.hadoop.hive.ql.io.parquet.MapredParquetInputFormat	org.apache.hadoop.hive.ql.io.parquet.MapredParquetOutputFormat	org.apache.hadoop.hive.ql.io.parquet.serde.ParquetHiveSerDe	

Schema

Partitions

Indexes

Column statistics - new

Schema (11)

View and manage the table schema.

Edit schema as JSON

Edit schema

Filter schemas

#	Column name	Data type	Partition key	Comment
1	data_venda	string	-	Data de venda para o cliente
2	data_producao	string	-	Data de produção do produto
3	right_razao_social	string	-	Razão Social do cliente
4	right_local	string	-	Locl de realização da venda
5	right_estado	string	-	Estado de realização da venda
6	codigo	bigint	-	Código de identificação do produto
7	descricao	string	-	Descrição do produto
8	preco	bigint	-	Preço de venda do produto por Kg
9	quantidade	bigint	-	Quantidade vendida do produto
10	peso	bigint	-	Peso vendido do produto
11	id	bigint	-	ID de identificação do cliente

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:

Dados

Fonte de dados  
AwsDataCatalog

Banco de dados  
mvp-dados-processados

Tabelas e visões  
Criar

Filtrar tabelas e visões

Tabelas (1)  
vendas-consolidadas  
data\_venda string  
data\_producao string  
right\_razao\_social string  
right\_local string  
right\_estado string  
codigo bigint  
descricao string  
preco bigint  
quantidade bigint  
peso bigint  
id bigint

Visões (0)

1 SELECT \* FROM "mvp-dados-processados"."vendas-consolidadas" limit 10;

SQL Ln 1, Col 70

Executar novamente Explicar Cancelar Limpar Criar

Reutilize os resultados da consulta até 60 minutos atrás

Resultados da consulta Estatísticas da consulta

Concluído Tempo na fila: 62 ms Tempo de execução: 410 ms Dados verificados: 20.19 KB

Resultados (10) Copiar Baixar resultados

Linhas de pesquisa

#	data_venda	data_producao	right_razao_social	right_local	right_estado	codigo	descricao
9	17/06/2023	08/10/2022	LAJEDO SERVICOS DE FESTAS E RECEPCOES LTDA	Rio de Janeiro (RJ)	RJ	3	Ternua das Vertentes 1000g
10	19/05/2023	08/10/2022	LAJEDO SERVICOS DE FESTAS E RECEPCOES LTDA	Rio de Janeiro (RJ)	RJ	3	Ternua das Vertentes 1000g
3	30/07/2023	23/03/2023	ADORO QUEIJO LTDA	Rio de Janeiro (RJ)	RJ	3	Ternua das Vertentes 1000g

Respostas:

1. Quais produtos vendem mais em quantidade ?

Criada Querie **MaisVendQTY**

Dados

Fonte de dados  
AwsDataCatalog

Banco de dados  
mvp-dados-processados

Tabelas e visões  
Criar

Filtrar tabelas e visões

Tabelas (1)  
vendas-consolidadas

Visões (0)

MaisVendPES : X MaisVendQTY : X MaisVendVAL : X PRINCLIVALCOMP : X ESTMAIVEND : X MESMAIVENDVAL : X

1 SELECT descricao,sum(quantidade) As QTY FROM "mvp-dados-processados"."vendas-consolidadas" group by (descricao) order by QTY desc;

SQL Ln 1, Col 130

Executar novamente Explicar Cancelar Limpar Criar

Reutilize os resultados da consulta até 60 minutos atrás

Resultados da consulta Estatísticas da consulta

Concluído Tempo na fila: 64 ms Tempo de execução: 490 ms Dados verificados: 6.52 KB

Resultados (32) Copiar Baixar resultados

Linhas de pesquisa

#	descricao	QTY
1	Hallounmi	3229
2	Be doce de leite	1858
3	Ternua das Vertentes 250g	1590



2. Quais produtos vendem mais em peso ?

Criada Querie **MaisVendPES**

Dados

Fonte de dados  
AwsDataCatalog

Banco de dados  
mvp-dados-processados

Tabelas e visões  
vendas-consolidadas

1

SELECT descricao, round(sum(peso/1000.0),2) As PESO\_KG FROM "mvp-dados-processados"."vendas-consolidadas" group by (descricao) order by PESO\_KG desc;

Executar novamente

Explicar

Cancelar

Limpar

Criar

Reutilize os resultados da consulta até 60 minutos atrás

Resultados da consulta

Estadísticas da consulta

Concluído

Tempo na fila: 58 ms

Tempo de execução: 425 ms

Dados verificados: 8.60 KB

Resultados (3/2)

Copiar

Baixar resultados

Linhas de pesquisa

#	descricao	PESO_KG
1	Be doce de leite	465.37
2	Massa de Boursin	439.48
3	Halloumi	375.18
4	Ternua das Vertentes 250g	364.66
5	Campones 250g	350.66

3. Quais produtos vendem mais em valor ?

Criada Querie **MaisVendVAL**

Dados

Fonte de dados  
AwsDataCatalog

Banco de dados  
mvp-dados-processados

Tabelas e visões  
vendas-consolidadas

1

SELECT descricao, round(sum(preco\*(peso/1000.0)),2) As VAL FROM "mvp-dados-processados"."vendas-consolidadas" group by (descricao) order by VAL desc;

Executar novamente

Explicar

Cancelar

Limpar

Criar

Reutilize os resultados da consulta até 60 minutos atrás

Resultados da consulta

Estadísticas da consulta

Concluído

Tempo na fila: 97 ms

Tempo de execução: 494 ms

Dados verificados: 10.57 KB

Resultados (3/2)

Copiar

Baixar resultados

Linhas de pesquisa

#	descricao	VAL
1	Ternua das Vertentes 250g	36465.6
2	Campones 250g	35066.3
3	Massa de Boursin	32960.85
4	Faixa de Carvao 1000g	31258.38
5	Halloumi	28138.35

4. Quais os principais clientes em valor de compras ?

Criada Querie **PRINCLIVALCOMP**

Dados

Fonte de dados  
AwsDataCatalog

Banco de dados  
mvp-dados-processados

Tabelas e visões  
Criar

▼ Tabelas (1)  
vendas-consolidadas  
data\_venda  
data\_producao  
right\_razao\_social  
right\_local  
right\_estado  
codigo  
descricao  
preco  
quantidade  
peso  
id

Visões (0)

1  
select ID, NOME, round(sum(VAL),2) AS VALOR from (select id, right\_razao\_social AS NOME, codigo, descricao, (peso/1000.0) AS PESO\_PROD, (preco\*(peso/1000.0)) AS VAL from "mvp-dados-processados"."vendas-consolidadas" order by id) group by (ID, NOME) order by VALOR DESC;

Executar novamente Explicar Cancelar Limpar Criar

Reutilize os resultados da consulta até 60 minutos atrás

Resultados da consulta Estatísticas da consulta

Concluído Tempo na fila: 61 ms Tempo de execução: 1.796 sec Dados verificados: 17.75 KB

Resultados (40)  
Copiar Baixar resultados

#	ID	NOME	VALOR
1	102	RIO DE JANEIRO	76808.35
2	40329989000104	LAJEDO SERVICOS DE FESTAS E RECEPCOES LTDA	51816.93
3	22686599000100	PARAMO AGROINDUSTRIA E COMERCIO LTDA	28642.5
4	101	OUTROS	14333.3
5	1	VENDA LOJA	13468.83

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

Criada Querie **ESTMAIVEND**

Dados

Fonte de dados  
AwsDataCatalog

Banco de dados  
mvp-dados-processados

Tabelas e visões  
Criar

▼ Tabelas (1)  
vendas-consolidadas  
data\_venda  
data\_producao  
right\_razao\_social  
right\_local  
right\_estado  
codigo  
descricao  
preco  
quantidade  
peso  
id

Visões (0)

1  
select Estado, round(sum(VAL),2) AS VALOR from (select right\_estado AS Estado, id, right\_razao\_social AS NOME, codigo, descricao, (peso/1000.0) AS PESO\_PROD, (preco\*(peso/1000.0)) AS VAL from "mvp-dados-processados"."vendas-consolidadas" order by right\_estado) group by (Estado) order by VALOR DESC;

Executar novamente Explicar Cancelar Limpar Criar

Reutilize os resultados da consulta até 60 minutos atrás

Resultados da consulta Estatísticas da consulta

Concluído Tempo na fila: 134 ms Tempo de execução: 502 ms Dados verificados: 9.38 KB

Resultados (7)  
Copiar Baixar resultados

#	Estado	VALOR
1	RJ	138430.98
2	MG	101172.23
3	SP	8440.25
4	SC	5655.42

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

Criada Querie **MESMAIVENDVAL**

The screenshot displays the AWS Athena interface. On the left, the 'Dados' sidebar shows the data source as 'AwsDataCatalog' and the database as 'mvp-dados-processados'. The main pane shows the SQL query for 'MESMAIVENDVAL'.

```
1 select MES, round(sum(VAL),2) AS VALMES From (SELECT data_venda, round(sum((preco*(peso/1000.0))),2) AS VAL, substring(data_venda,4,2) AS MES FROM "mvp-dados-processados"."vendas-consolidadas" group by (data_venda) order by data_venda, VAL desc) group by (MES) order by VALMES desc;
```

Below the query editor, the 'Executar novamente' button is visible. The 'Resultados da consulta' tab shows the query has completed successfully. The execution statistics are: Tempo na fila: 105 ms, Tempo de execução: 487 ms, Dados verificados: 13.72 KB.

The results table contains 12 rows, with the top 5 rows displayed:

#	MES	VALMES
1	05	37170.37
2	07	36945.13
3	06	34721.86
4	04	28420.34
5	08	25638.09

O objetivo proposto foi baseado em descobertas para definição de estratégia 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.