

MVP CONCLUSÃO DE SPRINT - CRIAÇÃO DE UM PIPELINE DE DADOS COM DELTA LIVE TABLES

—Lucas de Oliveira Noronha

Visão geral

O notebook "vendas" prepara e organiza dados de vendas em um pipeline DLT. Inicia configurando o ambiente e copiando os dados necessários para o DBFS. Em seguida, define uma tabela de fatos de vendas e tres dimensões (clientes, categorias e filial) para análise, garantindo a qualidade dos dados com expectativas DLT.

A base de dados é uma amostra de tres meses de vendas de uma empresa

https://raw.githubusercontent.com/LDONoronha/data_engineering/main/vendas.csv

Objetivos

1. Construir um data lake s para armazenar dados de emissões vendas de uma empresa.
2. Com essa base busco responder questões referentes a tendências de vendas e compras como por exemplo : Top 10 dos clientes com maior valor de compras em um respectivo mês; Qual departamento os clientes top 10 compram?

Plataforma

A plataforma escolhida foi a Databricks conforme orientações dos nossos professores.

Coleta de dados

O dataset escolhido foi uma base de vendas da empresa onde trabalho. Subi essa base em formato csv no meu repositório do github.



Link:

https://raw.githubusercontent.com/LDONoronha/data_engineering/main/vendas.csv

Modelagem e Carga

Metadados

Explorador de Catálogos > lucasdeoliveira > vendas >



















fato_vendas_view  

Visão geral Dados de exemplo Detalhes Permissões Histórico Linhagem Insights

Definição

```
SELECT
  COD_CLIENTE AS COD_CLIENTE
  /* PK CLIENTE */
  DEPARTAMENTO AS DEPARTAMENTO
  ...e mais 17 linhas
```

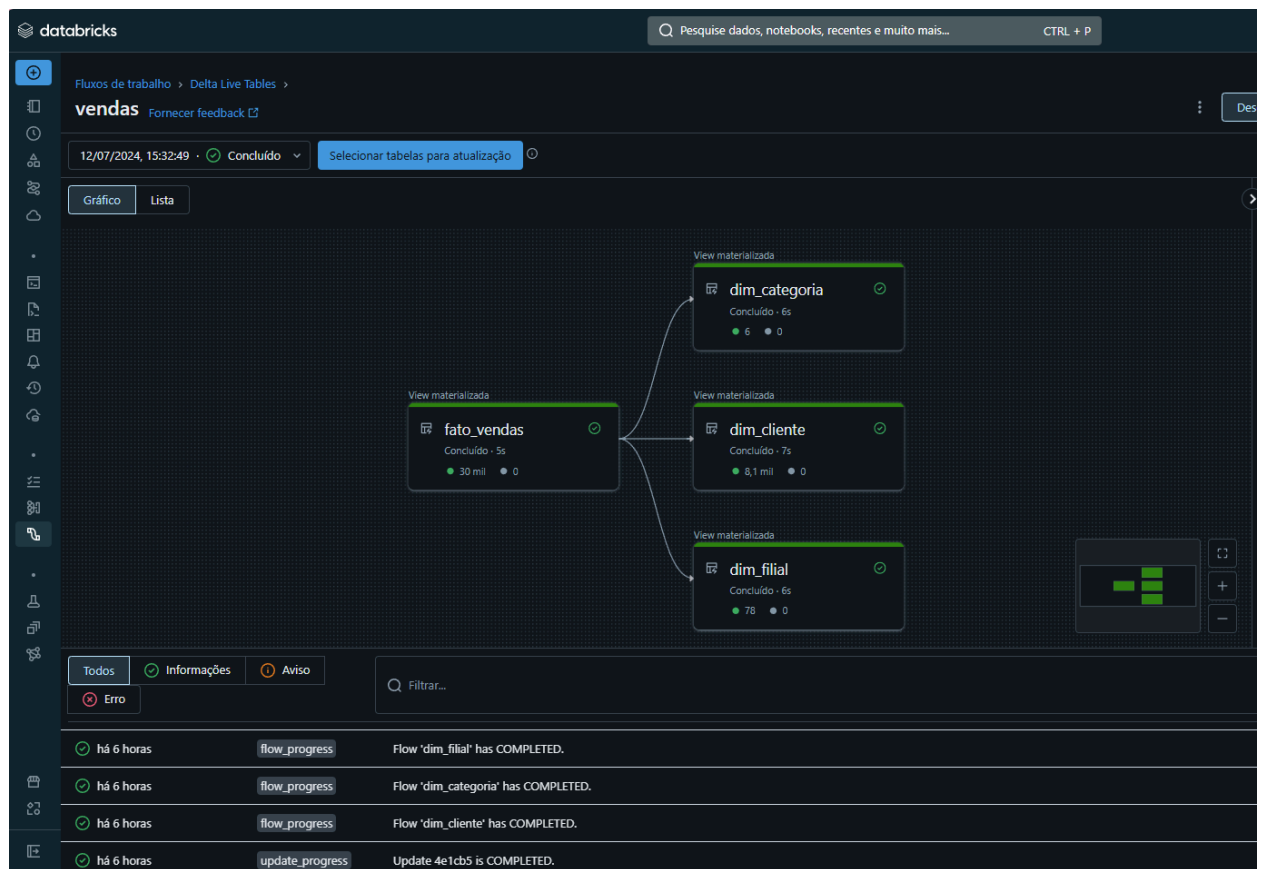
Q Filtrar colunas...

Coluna	Tipo	Comentário	Etiquetas	Regra de mascaramento de coluna
COD_CLIENTE	int	PK CLIENTE		
DEPARTAMENTO	string	Categoria ou família do produto		
MODO_PAGAMENTO	string	TIPO DE PAGAMENTO DO CLIENTE		
COD_FILIAL	int	LOJA		
UF	string	ESTADO		
QTDE	int	QUANTIDADE DE ITENS COMPRADOS		
VALOR_TOTAL	double	VALOR DE RECEITA DE VENDAS		
DATA	timestamp	DATA DE COMPRA		
MES_ANO	string	DATA DE COMPRA FORMATO RESUMIDO		

Linhagem dos dados



Esquema do Delta Live Tables



Agendamento dos jobs

The screenshot shows the Databricks Jobs interface for a job named 'atualizar_vendas'. The interface is in Portuguese. The top navigation bar includes the Databricks logo, a search bar, and a 'CTRL + P' shortcut. The left sidebar contains various icons for navigation. The main area displays the job configuration for 'atualizar_vendas' under the 'Tarefas' (Tasks) tab. A task named 'att_vendas' is shown, associated with the 'vendas' pipeline. Below the task list, there are fields for configuring the task: 'Nome da tarefa*' (Task name), 'Tipo*' (Type), 'Pipeline*' (Pipeline), 'Notificações' (Notifications), 'Novas tentativas' (New attempts), and 'Limite de duração' (Duration limit). The 'Tipo*' field is set to 'Pipeline das Delta Live Tables'. The 'Pipeline*' field is set to 'vendas'. There is a checkbox for 'Acionar uma atualização completa no pipeline Delta Live Tables' (Trigger a full update on the Delta Live Tables pipeline). The 'Notificações', 'Novas tentativas', and 'Limite de duração' fields have '+ Adicionar' (Add) buttons. At the bottom right, there are 'Cancelar' (Cancel) and 'Salvar tarefa' (Save task) buttons.

Fluxos de trabalho > Jobs > atualizar_vendas

Execuções Tarefas

att_vendas
Pipeline: vendas

+ Adicionar tarefa

Nome da tarefa* att_vendas

Tipo* Pipeline das Delta Live Tables

Pipeline* vendas

☐ Acionar uma atualização completa no pipeline Delta Live Tables

Notificações + Adicionar

Novas tentativas + Adicionar

Limite de duração + Adicionar

Cancelar Salvar tarefa

The screenshot shows the 'Agendas e triggers' (Schedules and triggers) section of the Databricks Jobs interface. The section is titled 'Agendas e triggers' and shows the schedule for the job: 'At 05:14 PM, on day 1 of the month (UTC-03:00 — Brasília)'. There are buttons for 'Editar trigger' (Edit trigger), 'Pausar' (Pause), and 'Deletar' (Delete). Below this, there is a section for 'Parâmetros do job' (Job parameters) with the text 'Nenhum parâmetro de job definido para este job' (No job parameters defined for this job) and a button for 'Editar parâmetros' (Edit parameters). There is also a section for 'Notificações do job' (Job notifications) with the text 'Sem notificações' (No notifications) and a button for 'Editar notificações' (Edit notifications). At the bottom, there is a section for 'Limites de duração' (Duration limits) with the text 'Nenhum limite definido' (No limit defined) and a button for 'Definir limites de duração' (Define duration limits).

Agendas e triggers

At 05:14 PM, on day 1 of the month (UTC-03:00 — Brasília)

Editar trigger Pausar Deletar

Parâmetros do job

Nenhum parâmetro de job definido para este job

Editar parâmetros

Notificações do job

Sem notificações

Editar notificações

Limites de duração

Nenhum limite definido

Definir limites de duração

Catálogo de dados

The screenshot shows the Databricks Catalog interface. On the left, the 'Catálogo' sidebar displays a tree structure of data sources. The 'vendas' schema is selected, showing a list of tables: 'dim_categoria', 'dim_categoria_view', 'dim_cliente', 'dim_filial', 'fato_vendas', and 'fato_vendas_view'. The main panel on the right, titled 'Explorador de Catálogos', shows the 'vendas' schema details, including a search bar for tables and a list of the same tables. The interface is dark-themed and includes a search bar at the top right.

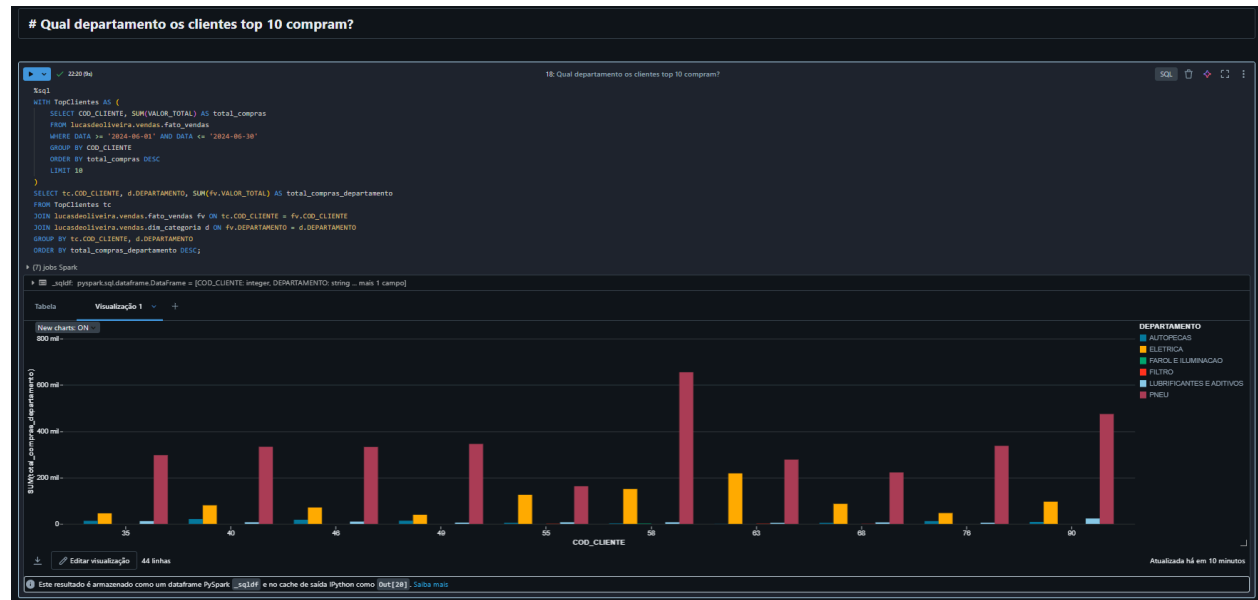
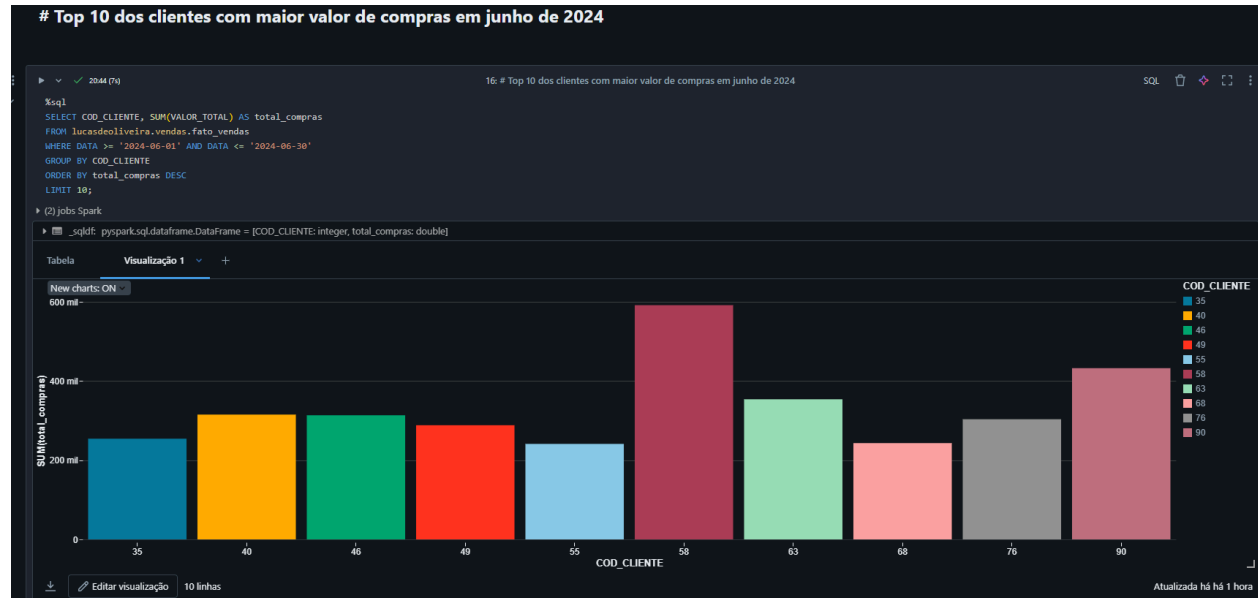
Análise

a. Qualidade de dados

No conjunto de dados não enfrentei problemas em relação a sua qualidade como com valores nulos e caracteres estranhos, pois o dataset estava bem limpo desde a sua origem.

The screenshot shows a CSV file with sales data. The header row is: `COD_CLIENTE,DEPARTAMENTO,MODO_PAGAMENTO,COD_FILIAL,UF,QTDE,VALOR_TOTAL,DATA,MES_ANO`. The data rows contain various sales records, including client codes, departments, payment methods, branch codes, states, quantities, total values, dates, and months. The data is displayed in a table format with a dark background.

b. Solução do problema



Autoavaliação

Esse projeto foi bem desafiador e bem empolgante de fazer, tive muitas dificuldades porém consegui atingir os objetivos e realizar o pipe line no databricks ajustando cotas e limites dos sistemas do google cloud. Sem dúvidas irei me aprofundar ainda mais nessa plataforma.