

## Planejamento do Trabalho

O objetivo deste trabalho é desenvolver uma análise detalhada dos dados de vendas da Amazon para identificar padrões de desempenho de diferentes categorias de produtos em diversos países. Este MVP deverá criar um pipeline de dados capaz de ingerir, processar, e analisar os dados.

## Objetivo

O objetivo deste MVP é identificar quais categorias de produtos estão se destacando e quais estão ficando para trás em termos de vendas podendo ajudar a empresa a tomar decisões sobre estratégias de marketing, logística, e alocação de recursos.

Para alcançar o objetivo do trabalho, as seguintes perguntas de negócio serão respondidas:

- Quais são as categorias de produtos de alta e baixa performance em cada país?  
Detalhamento: Identificar as categorias de produtos que apresentam os maiores lucros em cada país, fornecendo insights sobre os produtos mais populares e rentáveis em diferentes mercados.

**1. Dataset:** <https://www.kaggle.com/datasets/mithilesh9/amazon-sales-data-analysis>

**2. Coleta:** Download dos dados e upload manual no Databricks

**3. Modelagem:** Para modelar os dados da base AmazonSales, foi utilizada uma abordagem de esquema estrela. Esta abordagem facilita a análise de dados e a criação de relatórios, uma vez que os dados são organizados em tabelas de fatos e dimensões.

## Tabela de Fatos

- **fact\_amazon\_sales:** Contém os dados das vendas (fatos) relacionados aos pedidos.

## Tabelas de Dimensões

- **dim\_date:** Contém informações sobre datas.
- **dim\_orders:** Contém informações sobre os produtos.
- **dim\_location:** Contém informações sobre as regiões.

## 3.1 Catálogo de Dados

### Descrição Geral

Este catálogo de dados contém informações detalhadas sobre a base de dados AmazonSales, incluindo uma descrição de cada coluna, os tipos de dados, os valores mínimos e máximos esperados para dados numéricos e as possíveis categorias para dados categóricos.

## **Estrutura dos Dados**

Region

Country

Item type

Sales Channel

Order priority

OrderID

OrderDate

ShipDate

Units Sold

Unit Price

Unit Cost

Total Revenue

Total Cost

Total Profit

## **Descrição Detalhada dos Dados**

### **Region**

Descrição: Região de envio do pedido.

Tipo: Categórico

Valores Esperados: América do Norte, América Latina, América central, Europa, Ásia, África, Oceania

### **Country**

Descrição: País de envio do pedido.

Tipo: Categórico

Valores Esperados: Lista de países como Russia, Bulgaria, etc.

### **Item Type**

Descrição: Categoria do pedido.

Tipo: Categórico

Valores Esperados: Baby Food, Office Supplies, Household, etc.

### **Sales Channel**

Descrição: Canal de venda do pedido.

Tipo: Categórico

Valores Esperados: Online, Offline

**Order Priority**

Descrição: Prioridade do envio do pedido.

Tipo: Categórico

Valores Esperados: Alta (H), Média (M), Baixa (L), Crítica (C)

**OrderID**

Descrição: Identificador único do pedido.

Tipo: Numérico (Inteiro)

Valores Mínimos e Máximos: Valor numérico exclusivo para cada pedido.

**OrderDate**

Descrição: Data do pedido.

Tipo: Data

Formato: MM/DD/YYYY

Valores Esperados: Datas válidas no formato específico.

**ShipDate**

Descrição: Data de envio do pedido.

Tipo: Data

Formato: MM/DD/YYYY

Valores Esperados: Datas válidas no formato específico.

**Units Sold**

Descrição: Quantidade de produtos vendidos.

Tipo: Numérico (Inteiro)

Valores Mínimos e Máximos: Mínimo: 1, Máximo: dependendo do estoque disponível.

**Unit Price**

Descrição: Preço por unidade.

Tipo: Numérico (Decimal)

Valores Mínimos e Máximos: Mínimo: 0.01, Máximo: dependendo do tipo de produto.

**Unit Cost**

Descrição: Custo por unidade.

Tipo: Numérico (Decimal)

Valores Mínimos e Máximos: Mínimo: 0.01, Máximo: dependendo do tipo de produto.

**Total Revenue**

Descrição: Receita total.

Tipo: Numérico (Decimal)

Valores Mínimos e Máximos: Calculado como Units Sold \* Unit Price.

Total Cost

**Descrição: Custo total.**

Tipo: Numérico (Decimal)

Valores Mínimos e Máximos: Calculado como Units Sold \* Unit Cost.

**Total Profit**

Descrição: Lucro total.

Tipo: Numérico (Decimal)

Valores Mínimos e Máximos: Calculado como Total Revenue - Total Cost.

## 4. Carga

Carga feita no Databricks criando um cluster

Estado	Nome	Runtime	Memória ativa	Núcleos ativ...	DBU / h ativo	Origem	Criador	Notebooks
✓	MVP Eng dados Cluster (clone)	12.2	15 GB	2 núcleos	1	UI	biancalaragomes@h...	-
●	MVP Eng dados Cluster	12.2	-	-	-	UI	biancalaragomes@h...	-

Em seguida foi criado um notebook para a extração dos dados

Nome	Tipo	Proprietário	Criada às
ETL amazon sales 2024-07-09 22:50...	Notebook	bianca gomes	2024-07-09 22:50:45

Transformação dos dados feita em python podendo ser visualizado no github  
[https://github.com/biancalaragomes/MVP---ML-Analytics/blob/main/ETL%20amazon%20sales%202024-07-09%202022\\_50\\_44.py](https://github.com/biancalaragomes/MVP---ML-Analytics/blob/main/ETL%20amazon%20sales%202024-07-09%202022_50_44.py)  
Tabela carregada

ETL amazon sales 2024-07-09 22:50:44Python

ArquivoEditarVerExecutarAjudaA última edição foi em há 48 minutos

▶ Executar tudoMVP Eng dados Clust...CompartilharPublic

23:00 (2s)8SQL

```
select * from fact_amazon_sales
```

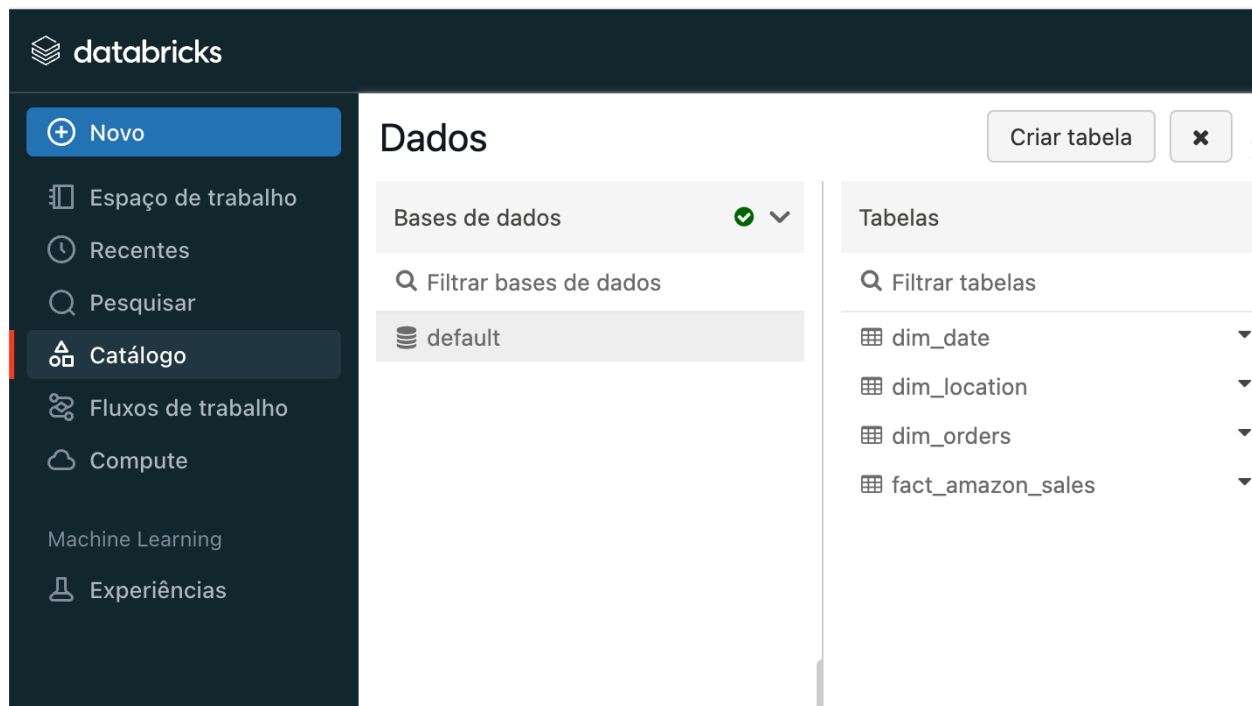
(2) jobs Spark

\_sqldf: pyspark.sql.dataframe.DataFrame = [region: string, Country: string ... mais 12 campos]

Tabela

	region	Country	Order_date	1.2 Unit_Price	1.2 Unit_Cost	1.2 Total_Revenue
1	Australia and Oceania	Tuvalu	2010-05-28	255.28	159.42	25
2	Central America and the Caribbe...	Grenada	2012-08-22	205.7	117.11	57
3	Europe	Russia	2014-05-02	651.21	524.96	1158
4	Sub-Saharan Africa	Sao Tome and Principe	2014-06-20	9.33	6.92	75
5	Sub-Saharan Africa	Rwanda	2013-02-01	651.21	524.96	3296
6	Australia and Oceania	Solomon Islands	2015-02-04	255.28	159.42	759
7	Sub-Saharan Africa	Angola	2011-04-23	668.27	502.54	2798
8	Sub-Saharan Africa	Burkina Faso	2012-07-17	154.06	90.93	1245
9	Sub-Saharan Africa	Republic of the Congo	2015-07-14	81.73	56.67	49
10	Sub-Saharan Africa	Senegal	2014-04-18	205.7	117.11	135
11	Asia	Kyrgyzstan	2011-06-24	154.06	90.93	19
12	Sub-Saharan Africa	Cape Verde	2014-08-02	109.28	35.84	455
13	Asia	Bangladesh	2017-01-13	109.28	35.84	902
14	Central America and the Caribbe...	Honduras	2017-02-08	668.27	502.54	5997
15	Asia	Mongolia	2014-02-19	81.73	56.67	400

Tabelas:



**Solução** – identificando categorias de produtos de alta e baixa performance por país

### Alta performance

O screenshot abaixo nos mostra que a categoria *cosméticos* é a mais lucrativa em 4 dos 5 países trazidos na consulta.

▶ Agora mesmo (1s) 7 SQL [ ] ⋮

```
%sql

select Item_Type, Country, sum(Units_Sold), sum(Total_Profit) as total_profit from fact_amazon_sales

group by Item_Type, Country
order by total_profit desc
limit 5
```

▶ (2) jobs Spark

▶ \_sqldf: pyspark.sql.dataframe.DataFrame = [Item\_Type: string, Country: string ... mais 2 campos]

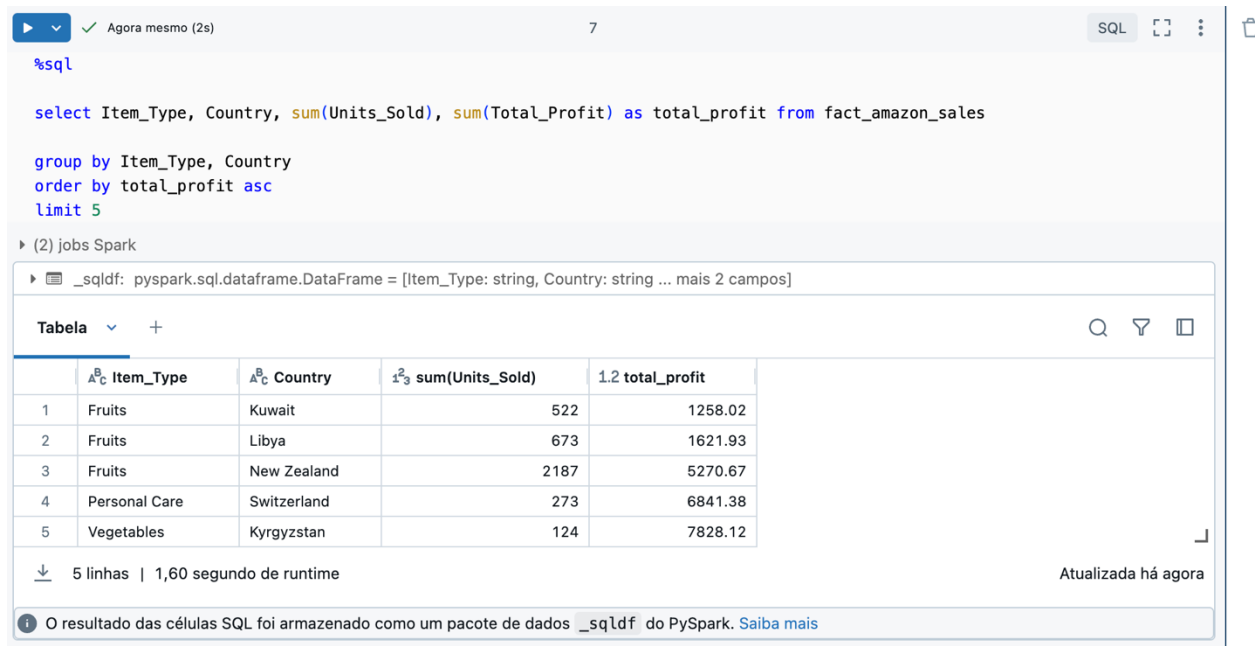
Tabela ▾ + 🔍 ⚙️ □

	Item_Type	Country	sum(Units_Sold)	total_profit
1	Cosmetics	Pakistan	9892	1719922.04
2	Cosmetics	Samoa	9654	1678540.98
3	Cosmetics	Iceland	8867	1541705.29
4	Cosmetics	Switzerland	8661	1505888.07
5	Household	Honduras	8974	1487261.02

↓ 5 linhas | 1,31 segundo de runtime Atualizada há agora

## Baixa performance

É possível visualizar que a categoria *Frutas* em 3 dos 5 países foi a que não teve uma boa performance em questão de lucro, mas isso se deve por conta do custo unitário ser baixo comparado à cosméticos.



The screenshot shows a Databricks SQL interface. At the top, there's a status bar with a play button, a checkmark, the text 'Agora mesmo (2s)', and a tab labeled '7'. To the right are buttons for 'SQL', a full-screen icon, and a menu icon. Below this is a code editor with the following SQL query:

```
%sql

select Item_Type, Country, sum(Units_Sold), sum(Total_Profit) as total_profit from fact_amazon_sales

group by Item_Type, Country
order by total_profit asc
limit 5
```

Below the code editor, it says '(2) jobs Spark'. Underneath that, a message indicates the query was saved: 'pyspark.sql.dataframe.DataFrame = [Item\_Type: string, Country: string ... mais 2 campos]'. The main part of the interface shows a table with the following data:

	Item_Type	Country	sum(Units_Sold)	total_profit
1	Fruits	Kuwait	522	1258.02
2	Fruits	Libya	673	1621.93
3	Fruits	New Zealand	2187	5270.67
4	Personal Care	Switzerland	273	6841.38
5	Vegetables	Kyrgyzstan	124	7828.12

At the bottom of the table, it says '5 linhas | 1,60 segundo de runtime'. To the right, it says 'Atualizada há agora'. At the very bottom, there's a message: 'O resultado das células SQL foi armazenado como um pacote de dados \_sqldf do PySpark. Saiba mais'.

## Autoavaliação

Objetivos alcançados foram

- **Modelagem de dados:** Foi criada uma estrutura de dados em esquema estrela, separando as dimensões e a tabela de fatos.
- **Ingestão de dados:** Os dados foram carregados no Databricks.
- **Catálogo de dados:** Um catálogo de dados foi desenvolvido, contendo descrições detalhadas dos campos, incluindo valores mínimos e máximos esperados para dados numéricos e categorias para dados categóricos.

Durante a execução do trabalho, algumas dificuldades foram encontradas:

1. **Erros de sintaxe:** Houve problemas iniciais com erros de sintaxe ao renomear colunas e construir o pipeline de dados.
2. **Problemas com caracteres inválidos:** Encontrei problemas ao tentar persistir dados com caracteres inválidos nos nomes das colunas.

3. **Complexidade de ferramentas:** Trabalhar com ferramentas avançadas como Databricks e Spark exigiu uma curva de aprendizado significativa.

## Trabalhos Futuros

Para enriquecer o problema e sua solução, os seguintes trabalhos futuros são recomendados:

1. **Monitoramento:** Implementar monitoramento para detectar problemas de qualidade dos dados.
2. **Análise avançada:** Integrar análises avançadas, como análise preditiva e machine learning.
3. **Visualização de dados:** Criar dashboards interativos e relatórios para visualização de dados, facilitando a interpretação dos resultados da análise.
4. **Integração com outras fontes de dados:** Expandir o pipeline para integrar dados de outras fontes.