

|  |  |  |
| --- | --- | --- |
|  |  | Page / |

|  |  |  |  |
| --- | --- | --- | --- |
| **PUBLIC** | INTERNAL | RESTRICTED | CONFIDENTIAL |

****

**Engenheiro de Soluções Júnior/Trainee**

Case Técnico

Olá, pessoa candidata, estamos muito ansiosos por ver mais de suas habilidades! Vamos lá?

# Objetivo

O objetivo deste teste é avaliar sua proficiência em alguns dos requisitos básicos para nossa vaga, que compreendem:

* Ensino superior completo nas áreas de Computação, Estatística, Matemática, Engenharias ou áreas afins;
* Conhecimento básico em Programação, preferencialmente SQL e/ou Python;
* Conhecimento em Estatística;
* Conhecimento em Machine Learning;
* Conhecimento em Visualização de Dados;
* Conhecimentos em Linux, Cloud e Metodologias Ágeis são um plus;
* Inglês Avançado;

Sobre a Dadosfera

* 2018: os fundadores, Luis Martins e Allan Sene, compartilhavam de uma mesma ideia revolucionar o mercado de dados no Brasil. Com isso, Luis descobriu uma forma de juntos, transformar a paixão por Dados em um negócio de sucesso orientado a Dados.
* 2019: No dia 20 de fevereiro de 2019 nasce, oficialmente, a empresa DataSprints.
* 2020: Começamos uma das maiores e mais renomadas consultorias de Data, AI & Analytics do Brasil.
* 2021: Recebemos nosso primeiro aporte de fundos de um investidor renomado no mercado com mais de 1 bilhão de dólares sob cessão! Lançamos nosso produto no mercado: Dadosfera e, estrategicamente, privatizamos a empresa com o mesmo nome.
* 2022: Conquistamos clientes e lançamos o nosso primeiro curso com a Dadosfera: Data Mesh na Prática!

# Sobre o Case

Nesse case entendemos não só suas habilidades para suportar o Comercial em Demos e Provas de Conceito, mas também, de estruturação da área de Solutions Engineering da Dadosfera, que está em franco desenvolvimento. Para isso, é essencial que você equilibre performance nos projetos de **estruturação de área**, expertise sobre **nossa plataforma** e também amplo **conhecimento na área de Dados**.

### Instruções:

* Em hipotese alguma compartilhe esse usuário com outra pessoa
* O uso de GPT e outros LLMs é **incentivado para esse case**. Faz parte do nosso dia-a-dia trabalhar com IA. Use com **sabedoria** para acertar todos os items, mas também sair **bem na entrevista técnica**.
* Seu acesso foi criado utilizando o **email de cadastro** **para esta vaga**. Voce pode achar o email pesquisando por "from: suporte@dadosfera.ai" ou o assunto "Acesse a Dadosfera | Registro":



* Siga [estas instruções](https://docs.dadosfera.ai/docs/primeiro-acesso) para realizar o seu primeiro acesso.

### Formato de Entrega:

Subir todas as respostas abaixo no Github, em um repositório com o nome <primeiro\_nome>\_<ultimo\_nome>\_DDF\_SOLUTIONS\_<mes><ano>, exemplo ALLAN\_OLIVEIRA\_DDF\_SOLUTIONS\_082023. Responder os itens em formato **Markdown** com **LINKS direcionando para os ativos na Dadosfera** e **PRINTS** evidenciando os datasets, dashboards e data app.

Para o Item 1:

1 - O video deve ser carregado como **Unlisted** no **YouTube**. Seu link deve ser referenciado no **markdown** acima.

2 - Garanta que o video é acessível por **qualquer usuário** com o link.

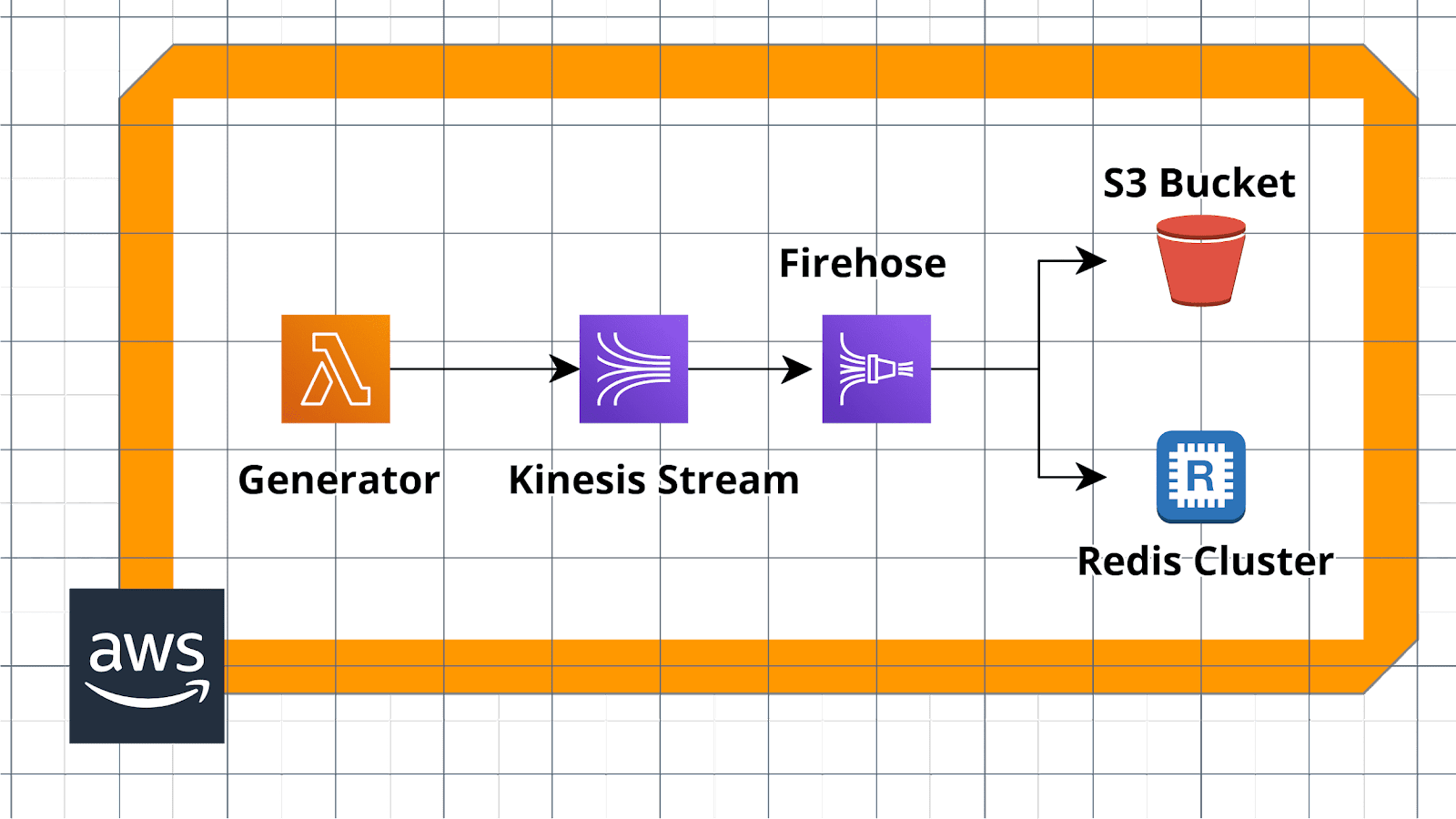
3 - Qualquer outra análise ou trabalho feito **fora da Dadosfera**, deve ser explicada no video e estar **listada e linkada claramente** na Dadosfera e na Apresentação

Requisitos:

* **Ter acesso à Internet**
* **Ter acesso ao ambiente de Treinamentos da Dadosfera**

# **Item 1 - Sobre Solutions Engineering**

Seu cliente, uma grande empresa de e-commerce, está com um projeto de um novo aplicativo de dados em produção. Após a primeira reunião com nossa equipe de vendas, você deve fornecer uma solução técnica que atenda às exigências e se encaixe na arquitetura deles. Aqui está o diagrama da arquitetura atual deles:



Você deve mostrar, com base nos seus estudos sobre a [Documentação da Dadosfera](http://docs.dadosfera.ai), a **viabilidade de substituirmos essa solução**.

Essa arquitetura deve ser a base de uma nova Plataforma de Dados, com o objetivo de fornecer:

* Modelos de IA para melhorar a experiência de compra dos clientes

Sua apresentação deve ter, pelo menos, 3 destas respostas:

* Principal problema a ser resolvido
* Diagrama da solução com Dadosfera, substituindo parcial ou totalmente a arquitetura desenhada acima.
* Como a Dadosfera é uma solução tecnicamente mais viável e/ou mais barata
* Oportunidades e ganhos futuros de se adotar a Dadosfera, frente à solução atual

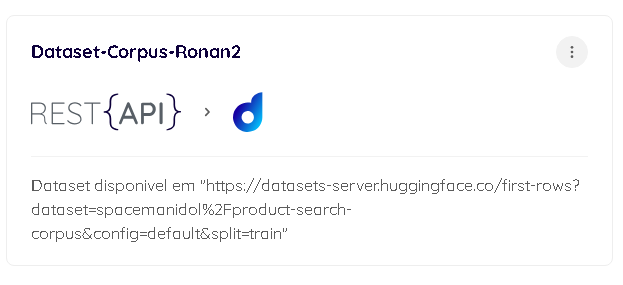
Seja claro na apresentação, explicando o seu pensamento e como deve funcionar em cenários reais. Lembre-se de se apresentar como um Engenheiro de Soluções da Dadosfera (pensamento positivo 🙏🏾).

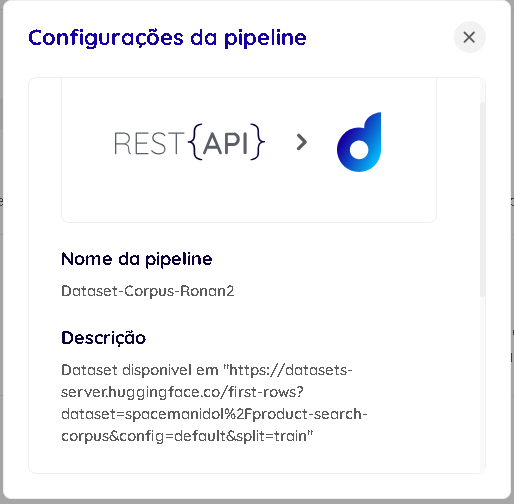
**Link do Video: https://youtu.be/o0vZe84xz\_c**

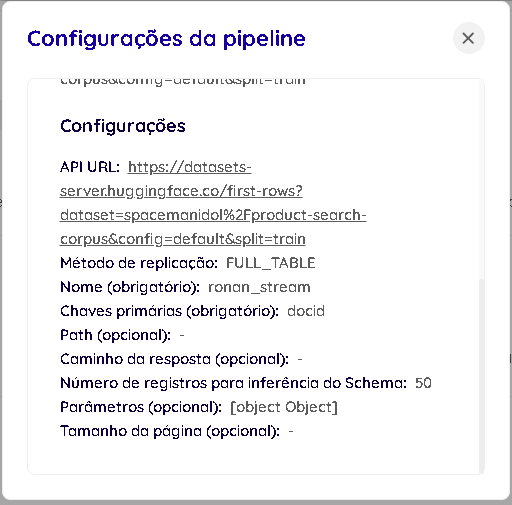
# Item 2 - Sobre a Dadosfera

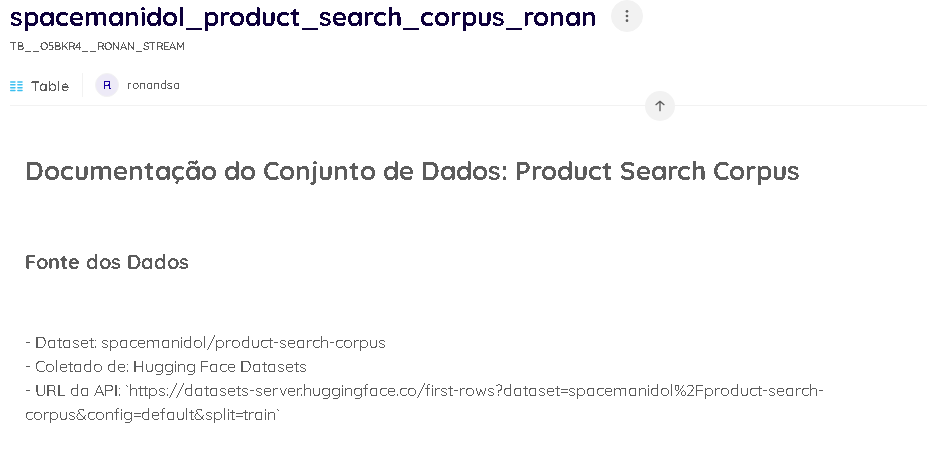
Parabéns! Conseguimos convencer o cliente de usar a Dadosfera. Agora, precisamos iniciar o projeto de implementação da nossa solução para um case específico. Vamos iniciar com o **carregamento** e posterior **análise descritiva** dos dados de produtos de um site para análise. Os dados estão disponíveis [neste site](https://huggingface.co/datasets/spacemanidol/product-search-corpus/viewer/default/train?row=0). Toda a documentação sobre o dataset também esta disponível.

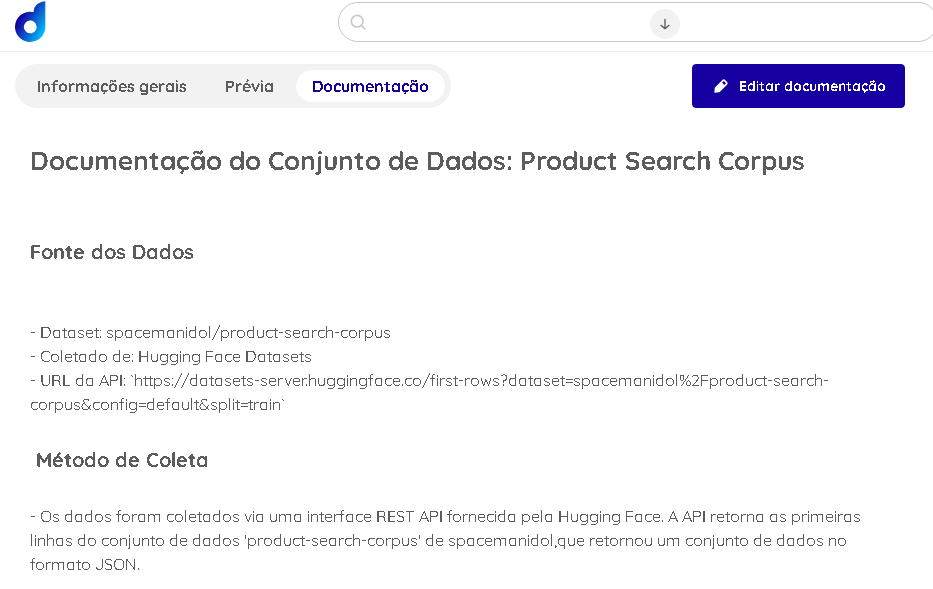
Usando os seus conhecimentos da documentação da Dadosfera, **faça a carga desse dataset**, **catalogue-o** com as informações mais relevantes, seguindo boas práticas de Dicionário de Dados.

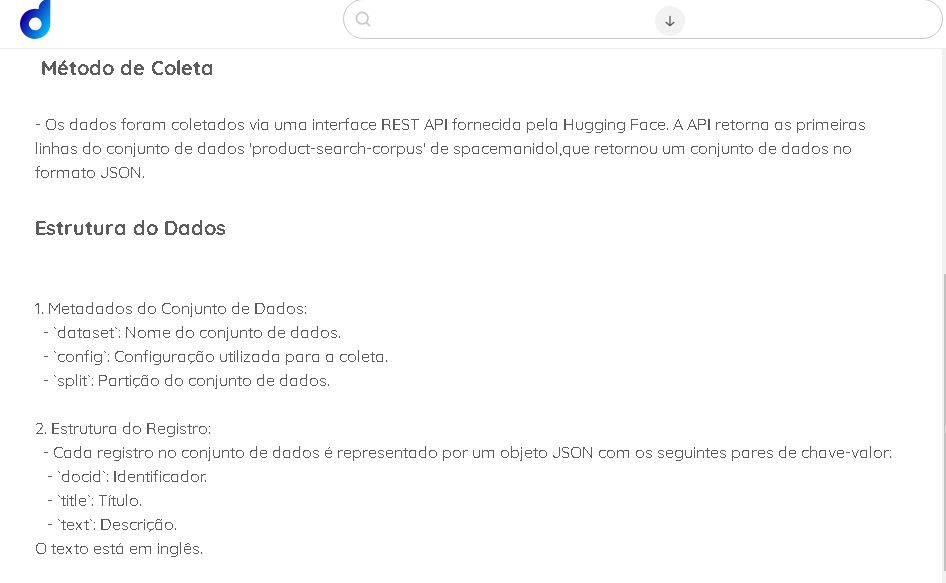


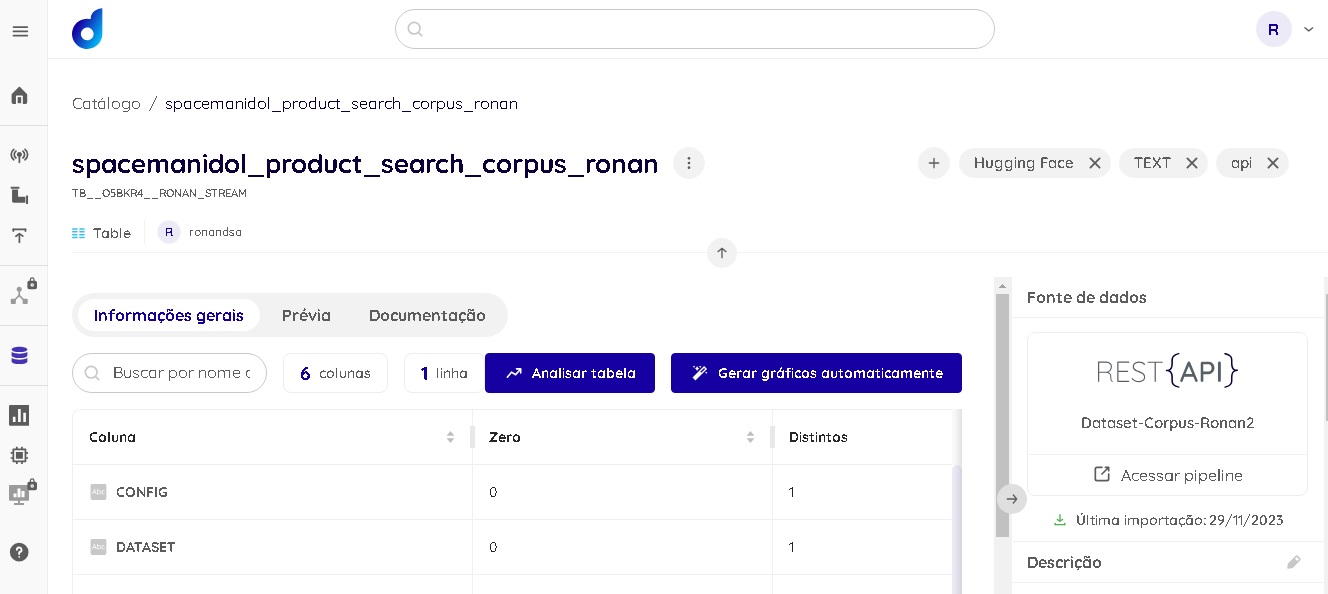












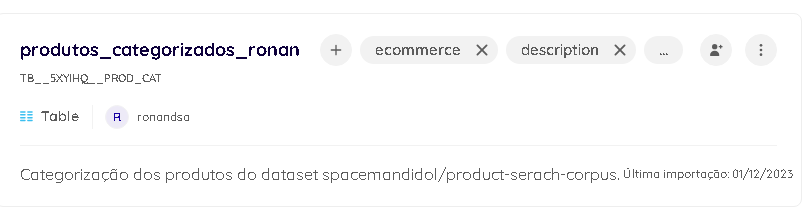
# 

# Item 3 - Sobre GenAI e LLMs

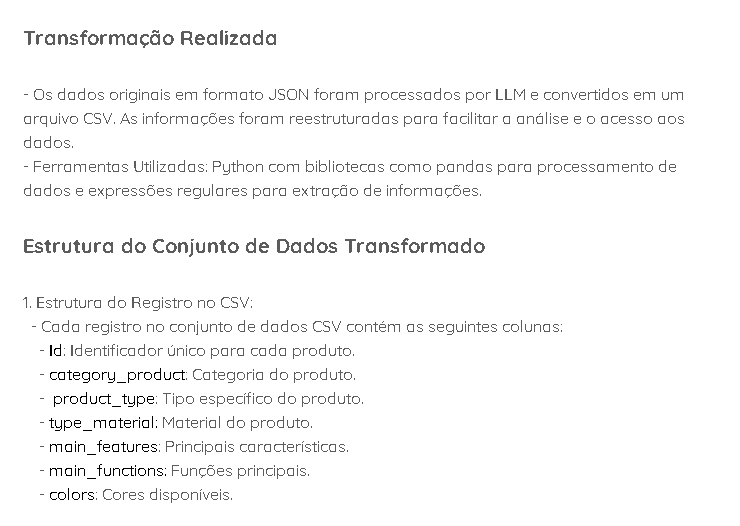
O Dataset disponibilizado tem 2 colunas de texto, uma de título e outra de descrição. Utilize um LLM - como ChatGPT ou Claude - para criar features dos produtos que serão úteis para análise. Exemplo:

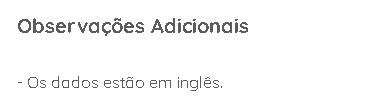
|  |
| --- |
| **Titulo**: FYY Leather Case with Mirror for Samsung Galaxy S8 Plus, Leather Wallet Flip Folio Case with Mirror and Wrist Strap for Samsung Galaxy S8 Plus Black |
| **Product Description**: Premium PU Leather Top quality. Made with Premium PU Leather. Receiver design. Accurate cut-out for receiver. Convenient to Answer the phone without open the case. Hand strap makes it easy to carry around. RFID Technique: Radio Frequency Identification technology, through radio signals to identify specific targets and to read and copy electronic data. Most Credit Cards, Debit Cards, ID Cards are set-in the RFID chip, the RFID reader can easily read the cards information within 10 feet(about 3m) without touching them. This case is designed to protect your cards information from stealing with blocking material of RFID shielding technology. 100% Handmade. Perfect craftsmanship and reinforced stitching make it even more durable. Sleek, practical, and elegant with a variety of dashing colors. Multiple Functions Card slots are designed for you to put your photo, debit card, credit card, or ID card while on the go. Unique design. Cosmetic Mirror inside made for your makeup and beauty. Perfect Viewing Angle. Kickstand function is convenient for movie-watching or video-chatting. Space amplification, convenient to unlock. Kickstand function is convenient for movie-watching or video-chatting. |
| **Features**: {  "category": "Phone Accessories",  "material": "Premium PU Leather",  "features": {  "receiver\_design": "Accurate cut-out for receiver. Convenient to Answer the phone without opening the case.",  "hand\_strap": "Yes",  "RFID\_technique": "Protection of card information with RFID shielding technology",  "handmade": "100% Handmade",  "stitching": "Reinforced stitching",  "functions": {  "card\_slots": "Yes",  "cosmetic\_mirror": "Yes",  "kickstand\_function": "Yes, convenient for movie-watching or video-chatting",  "space\_amplification": "Yes, convenient to unlock"  },  "color\_options": "Variety of dashing colors",  "compatibility": "Samsung Galaxy S8 Plus"  }  } |

Para entender como usar um LLM como o GPT, acesse este notebook de exemplo ou então pesquise no Playground da Open AI ([referenciado abaixo](https://docs.google.com/document/u/5/d/e/2PACX-1vRMX5wK2mgyIQSTuKi4dGhjNPHyodE4Je2M-SU6ujnsq9JrpoBrX36JdallEzoLtrIVXC-OAO1KsdYf/pub?utm_medium=email&_hsmi=278157465&_hsenc=p2ANqtz-8uN2SK6fddj-bCcUc8LNvr604dBaOnKnvRxqlSiDMwCutgwxj-YKWol65IXrYUirpYmp1gyaQHP8Eay7ZbFECU3tfxww&utm_content=278157465&utm_source=hs_automation#h.9b2oakpnrkv3))





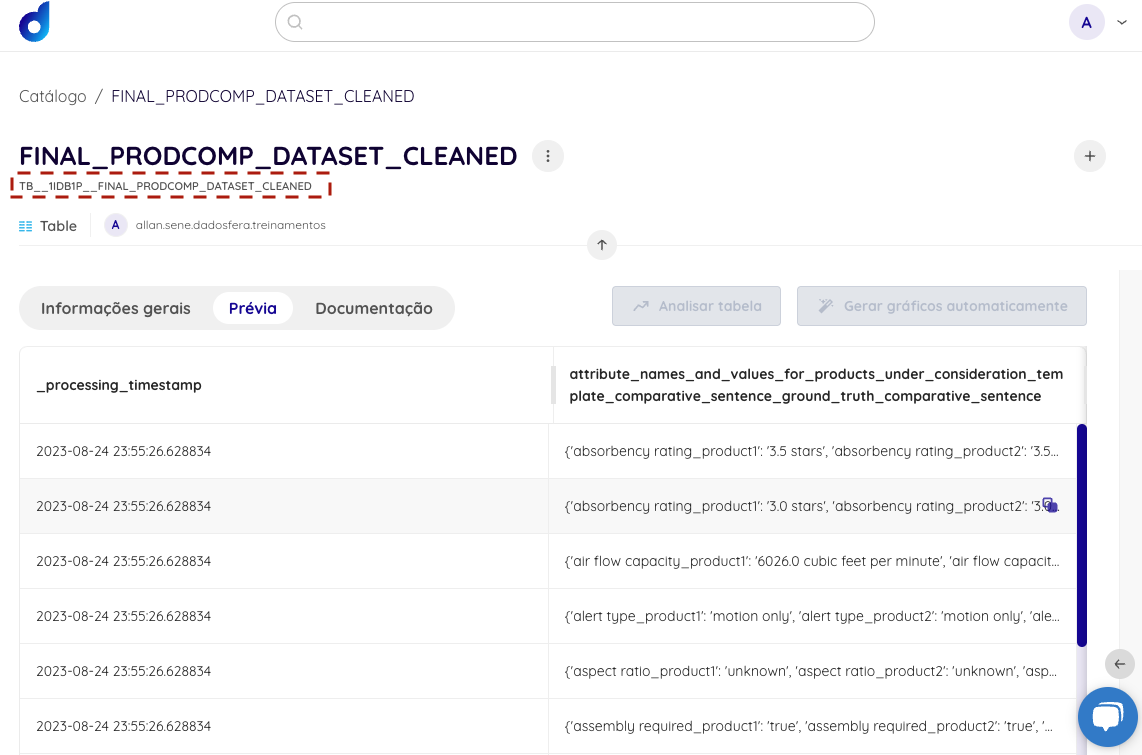




# Item 4 - Sobre SQL e Python

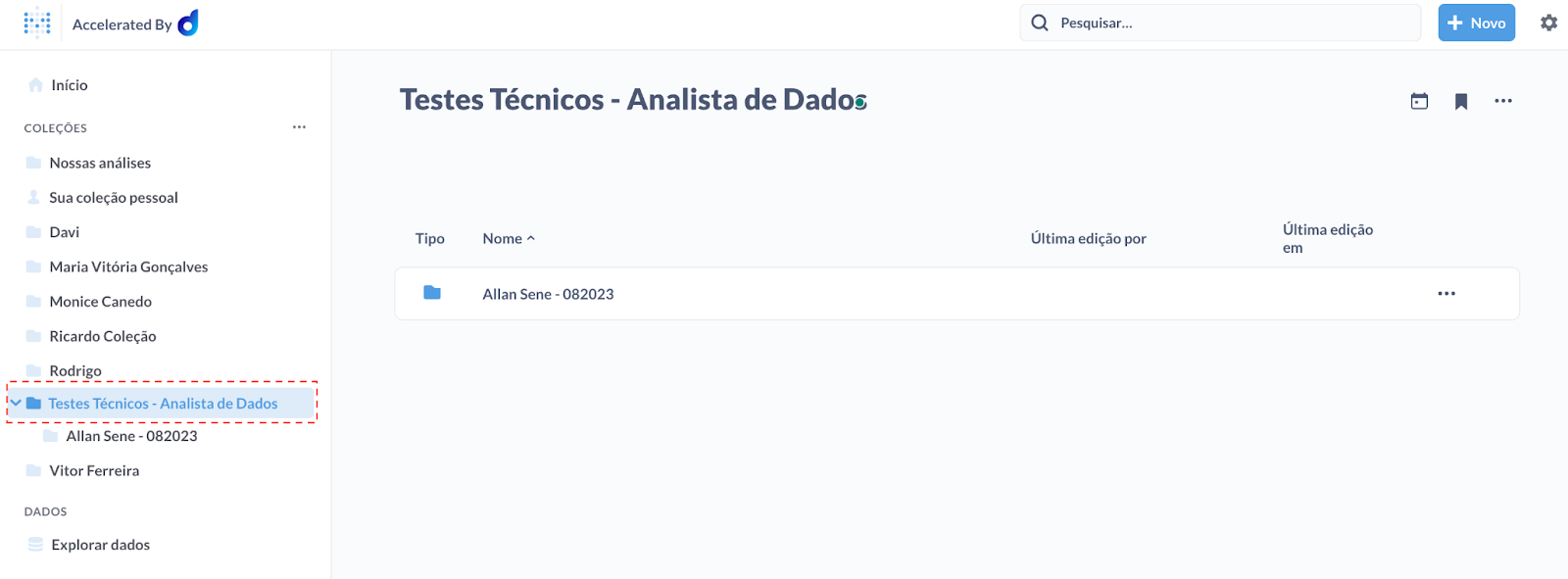
Com as features extraídas pelo LLM dentro da Dadosfera, precisamos ir para a etapa de geração de valor.

Utilize as mesmas credenciais para acessar nosso módulo de Visualização. Observe que você precisa utilizar o identificador (ID) da tabela para encontrá-la no nosso [módulo de Visualização](https://docs.dadosfera.ai/docs/visualiza%C3%A7%C3%A3o):

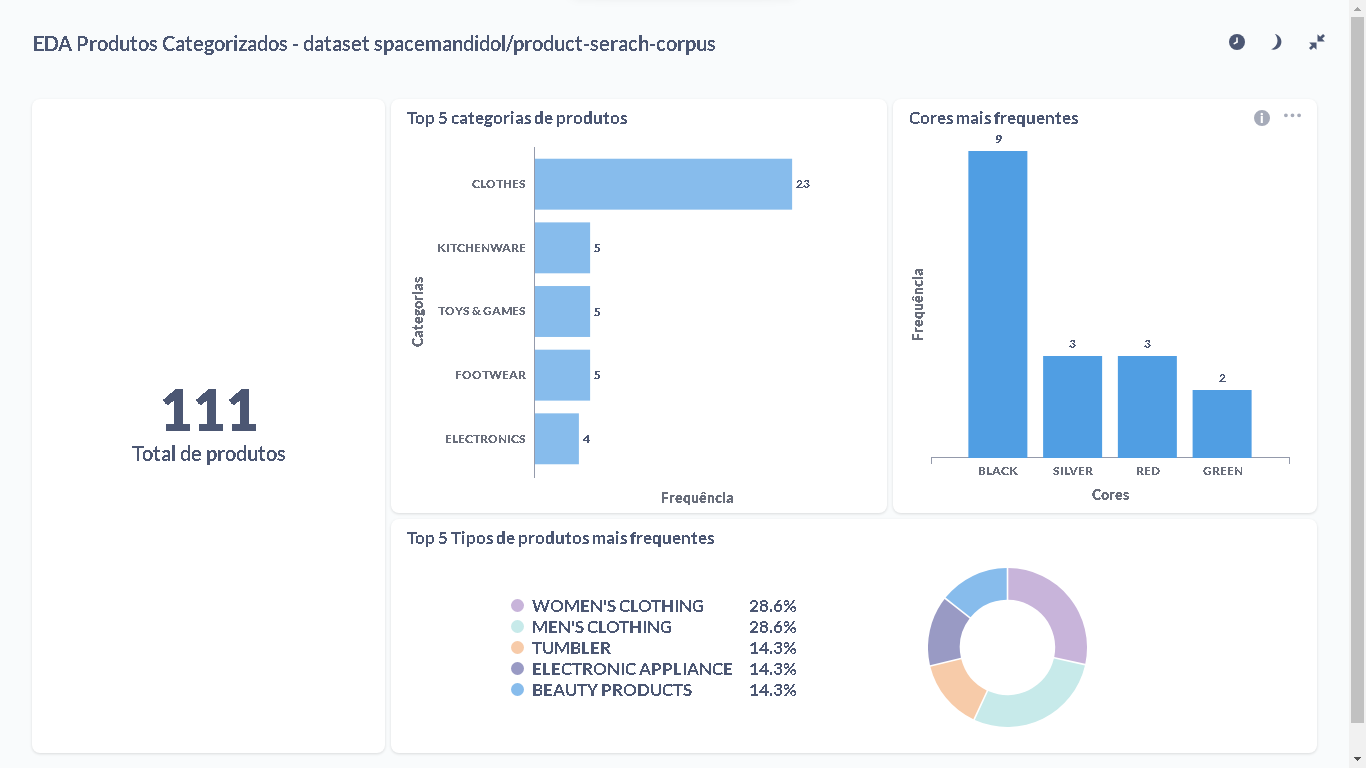


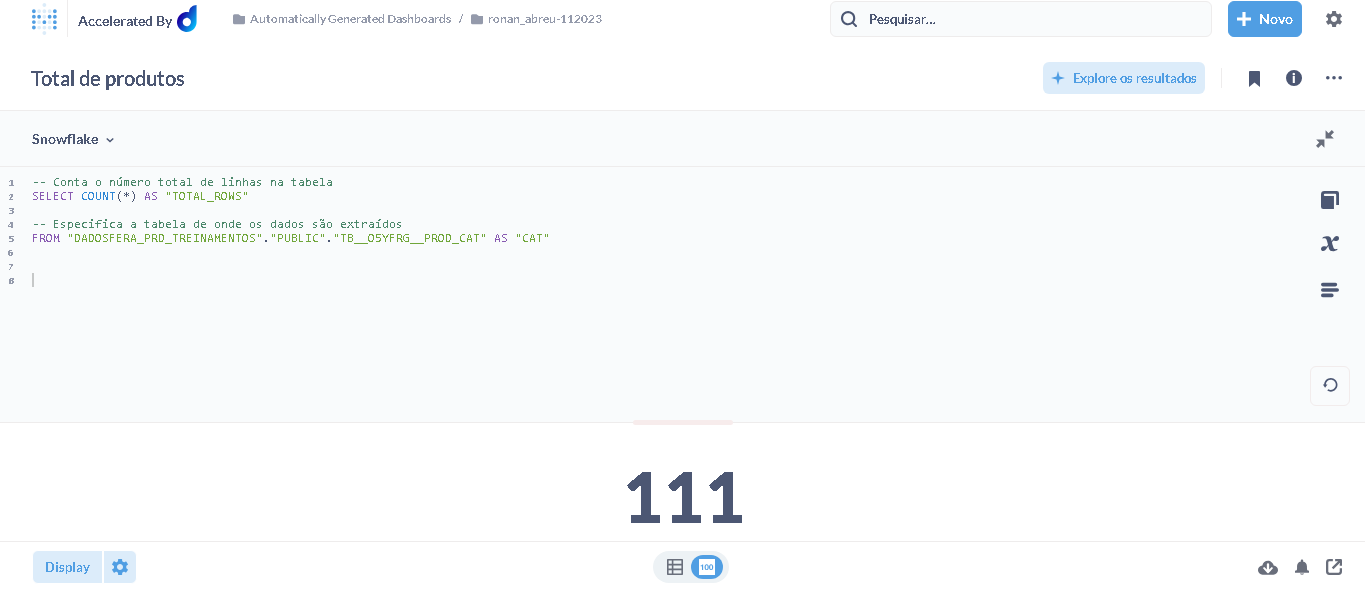
*Exemplo do dataset na Dadosfera, destacando o ID*

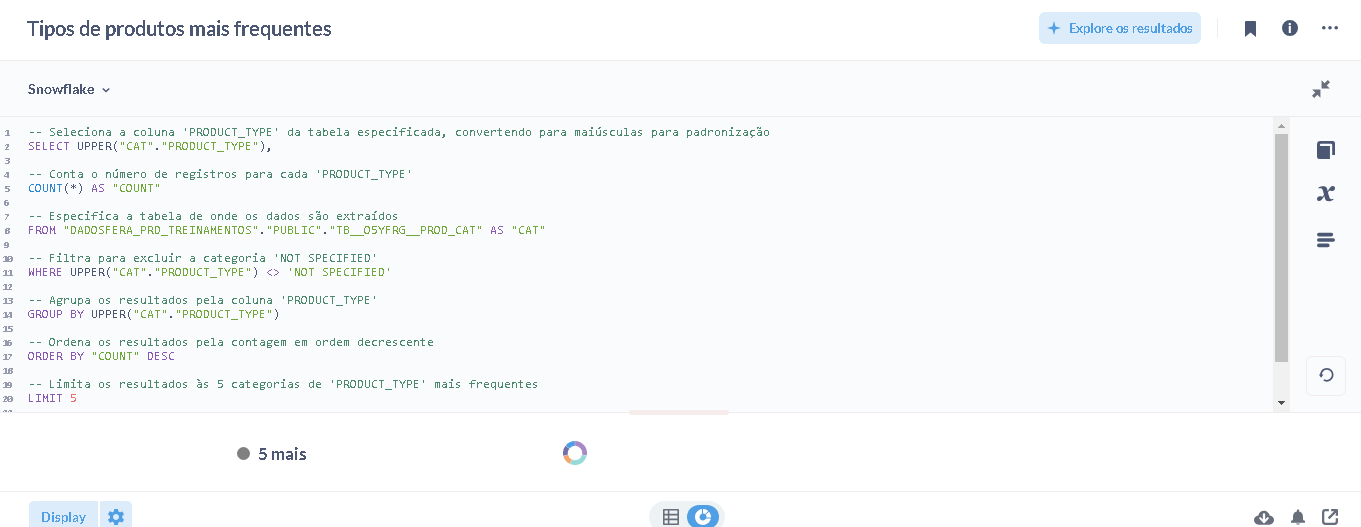
Crie uma Coleção com o formato <nome> <sobrenome> - <mes\_ano> assim como no exemplo:

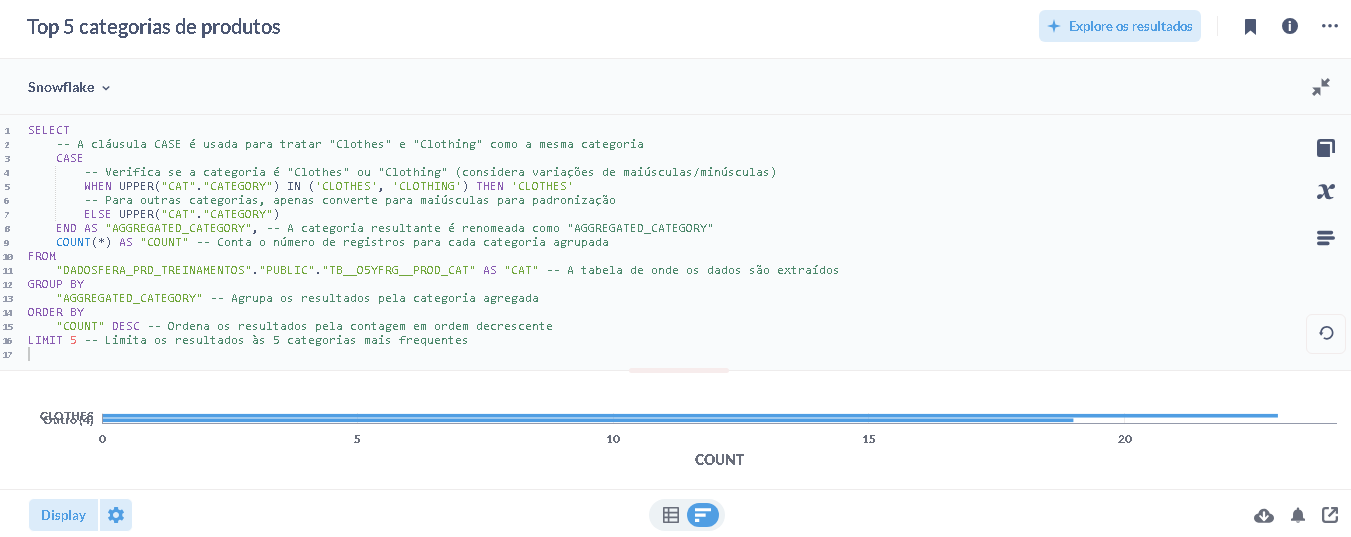


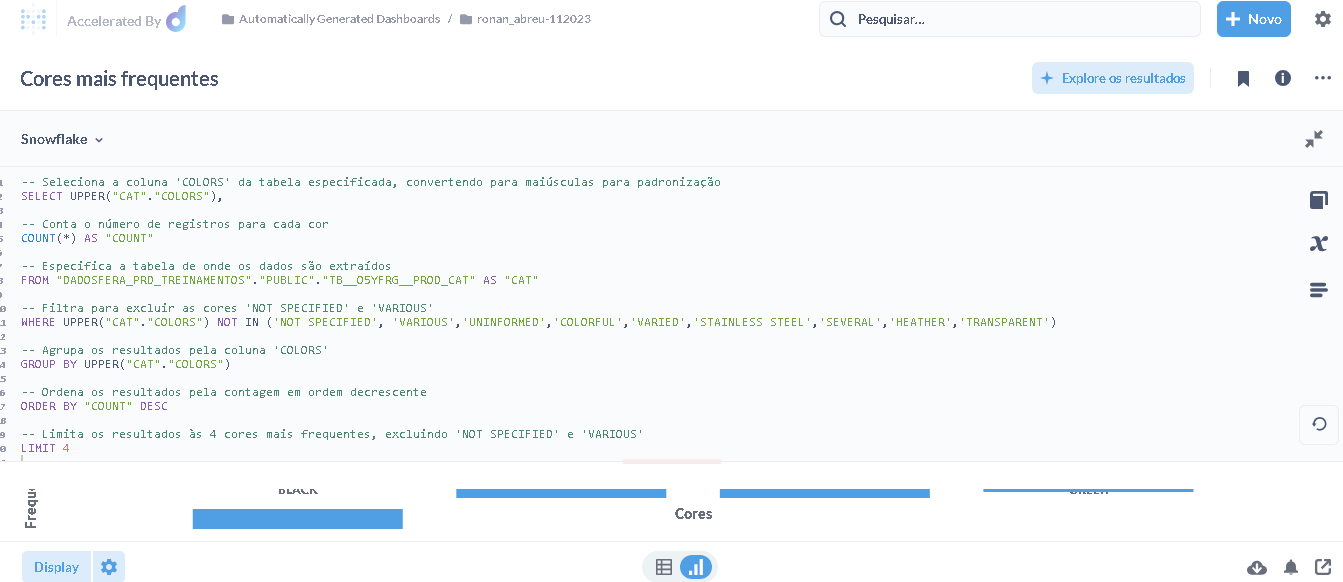
Crie uma visualização de dados que mostre uma **análise das categorias dos produtos**. Salve a Query SQL utilizada e também o print do resultado da query no documento markdown deste teste.











Interface gráfica do usuário, Texto, Aplicativo, Teams

Descrição gerada automaticamente

# Item 5 - Sobre Data Apps

Agora você tem que criar um Data App utilizando Streamlit para explorar os dados anteriores. Para criar um Data App, acesse nosso módulo de inteligência e siga [este guia na nossa documentação](https://docs.dadosfera.ai/docs/data-app-com-streamlit).

Sugestões de Análises:

* [Similaridade entre produtos](https://www.kaggle.com/code/jeffd23/visualizing-word-vectors-with-t-sne/notebook)
* [Tensorboard - Similaridade entre produtos](https://towardsdatascience.com/how-to-visualize-text-embeddings-with-tensorboard-47e07e3a12fb)
* [EDA - usando GPT](https://blog.streamlit.io/building-a-streamlit-and-scikit-learn-app-with-chatgpt/)

Texto

Descrição gerada automaticamente

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

<https://app-intelligence-treinamentos.dadosfera.ai/pbp-service-streamlit-50d9c6e8-bcc9-4e6143da4674-ca2d-45d3_8501/>

Tabela

Descrição gerada automaticamente

Texto

Descrição gerada automaticamente

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

# Item Bonus - Sobre GenAI + Data Apps

Utilize este Data App como referência para criar um [gerador de apresentações de produto](https://blog.streamlit.io/instant-insight-generate-data-driven-presentations-in-a-snap/). Seu objetivo é chegar em uma apresentação para mostrar as principais características do Produto a fim de vender mais. Utilize a [API do Dall-e](https://platform.openai.com/docs/api-reference/images) ou outro [gerador de imagens](https://huggingface.co/stablediffusionapi) para gerar uma imagem do produto. Exemplo de uso [do Dall-e aqui](https://app-intelligence-treinamentos.dadosfera.ai/file-preview?project_uuid=f2cda60b-041a-495c-b39f-4507126614bd&pipeline_uuid=dcd74f1c-ab3d-4f26-87bc-54d7a11b223a&step_uuid=54ab516b-b889-4879-a2dd-2444856983e0):



## Resultado e Forma de Avaliação

Em escala de nota final menor para a maior.

* **Mínimo**: responder as perguntas do item do case 2, 3 e 4 e carregá- no Github
* **Intermediário**: mínimo anterior + item do case 1 (apresentaçao em video)
* **Avançado**: Intermediário + item case 5 (Data App em streamlit)
* **Excelente**: Avançado + case bonus
* **Outlier**: extrapolar o que foi pedido.

## Referências para esse teste

* [**Documentação da Dadosfera**](https://docs.dadosfera.ai/docs)
* [**Galeria de Data Apps do Streamlit**](https://streamlit.io/gallery)
* [**Open AI Playground**](https://platform.openai.com/playground)
* [**API da Dadosfera**](https://docs.dadosfera.ai/reference)
* [**Fórum de Discussões da Dadosfera**](https://docs.dadosfera.ai/discuss)
* [**Documentação do Metabase**](https://www.metabase.com/docs/latest/)