## **Web Scraping com n8n**

### **1. Abertura e contexto (5 min)**

**Tópicos**:

* O que é Web Scraping e por que é uma habilidade essencial para automação e coleta de dados externos
* Exemplos reais de uso: monitoramento de licitações, inteligência de preços, coleta de leads B2B, análise de reputação
* Conexão com automações profissionais: Web Scraping como ponto de entrada de dados em pipelines

### **2. Conceitos fundamentais (10 min)**

**Tópicos**:

* Diferença entre scraping manual, APIs oficiais e plataformas intermediárias
* O que são APIs: explicação com analogia a pipelines e DataMarts
* Explicação prática: cada API como um SELECT na web
* Apify: estrutura baseada em Actor e Dataset, exposição via API REST
* Visualização com gráfico Mermaid do fluxo Actor > Dataset > API > Ferramentas (n8n, Python, Power BI)

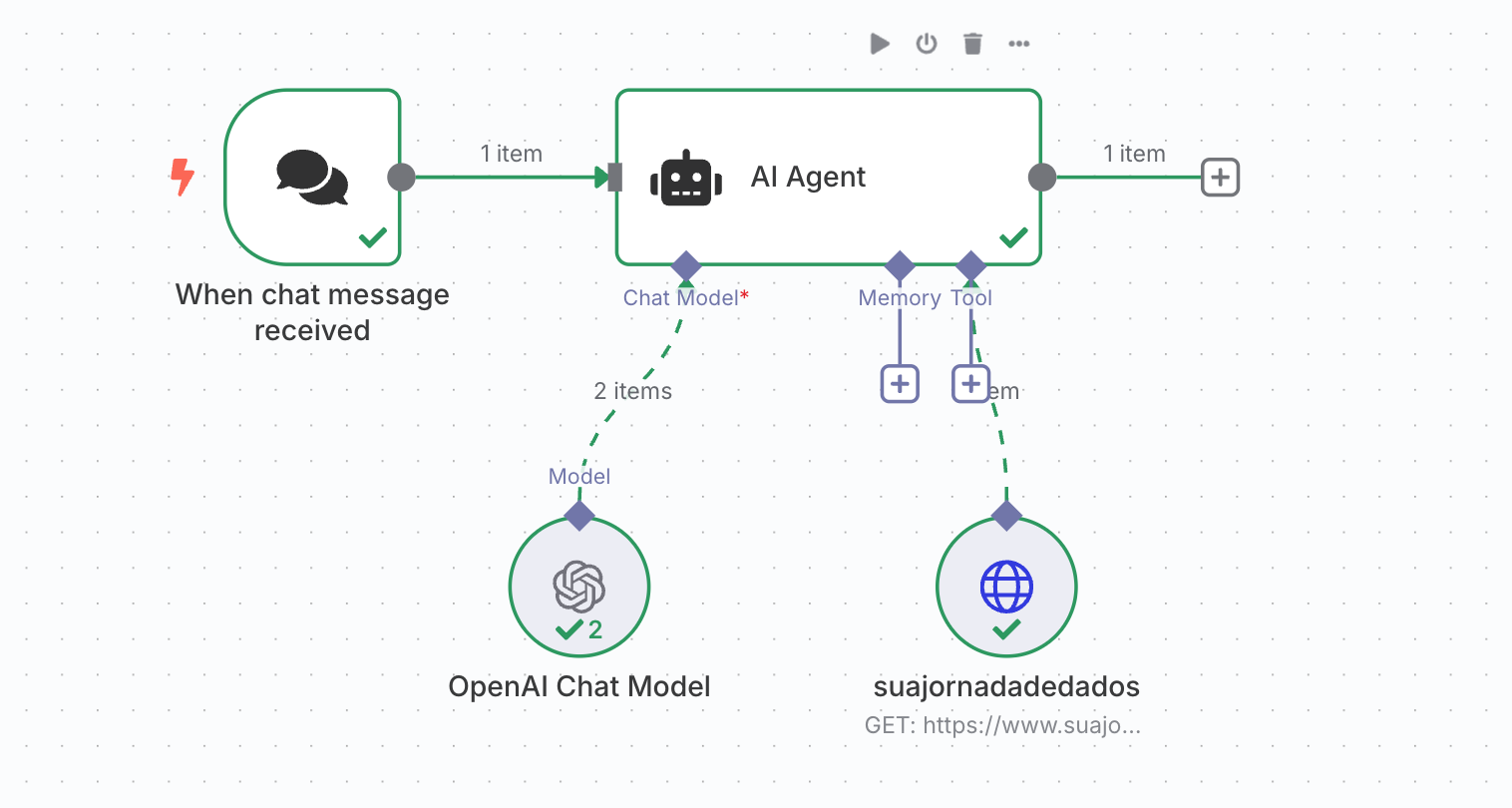
### **3. Abordagens práticas de Web Scraping com n8n**

#### **a) Agente com Web Scraping Direto do Site (25 minutos)**

### **Aplicação prática**

Esse prompt é ideal para:

* MVPs com scraping como backend
* Atendimento automatizado em sites próprios
* Chatbots de onboarding ou suporte informativo
* Versão inicial de agentes conectados a conteúdo público



### **Explicação dos nós utilizados**

#### **1. When Chat Message Received**

* **Captura a entrada do usuário.  
  Passa a mensagem via chatInput para o agente.**

**AI Agent**

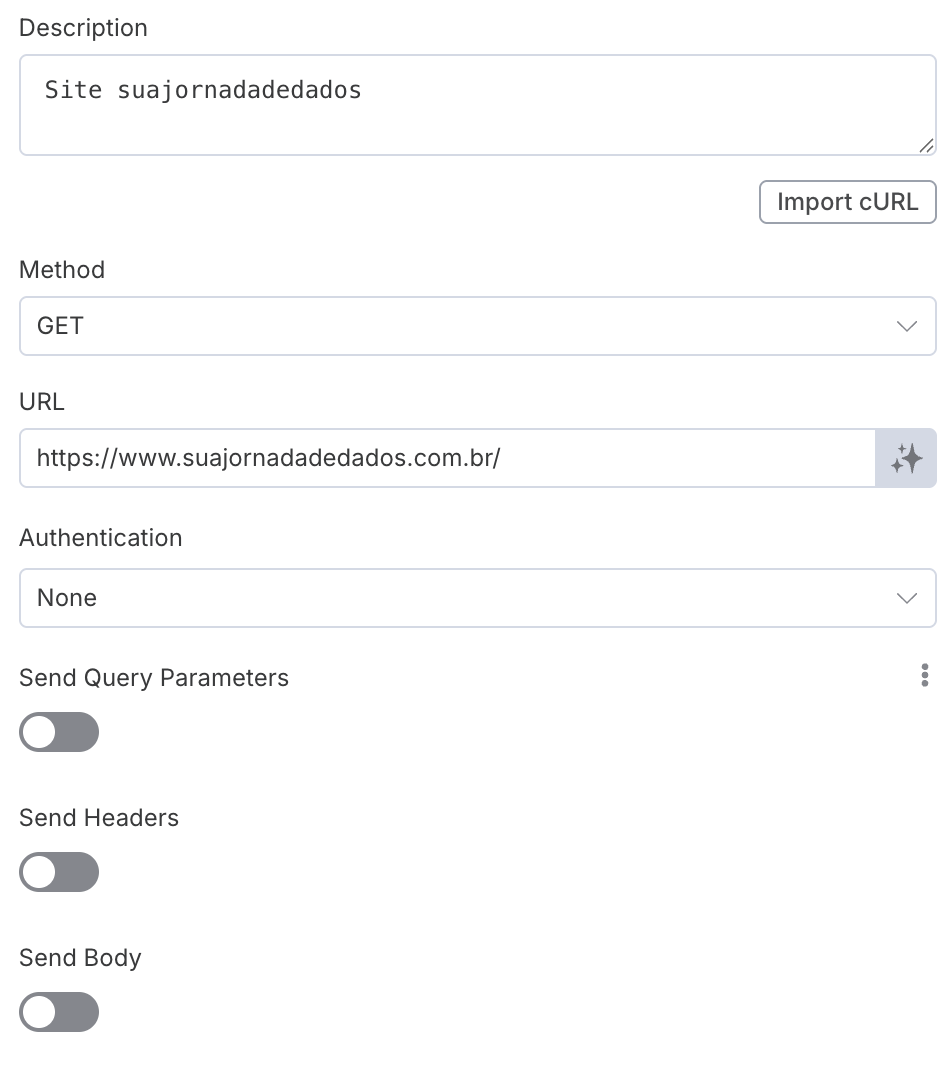
* **É o cérebro do fluxo.**
* **Usa o modelo da OpenAI para processar a pergunta e decidir se deve ou não ativar a ferramenta suajornadadedados.**
* **Está conectado a:**
  + **OpenAI Chat Model como modelo principal**
  + **suajornadadedados como única Tool disponível**

#### **3. OpenAI Chat Model**

* **Utiliza gpt-4.1 para raciocínio e decisão.**
* **Recebe a instrução do agente e retorna a resposta final (diretamente ou após usar a Tool).**

#### 

#### **4. suajornadadedados**



* **Realiza scraping direto no site www.suajornadadedados.com.**
* **Traz conteúdos como:**
  + **Programação de aulas**
  + **Workshops**
  + **Professores**
  + **Calendário ou links**
* **Responde em formato estruturado (JSON) para ser transformado em texto pelo agente.**

## 

## **System Message**

**Você é um agente inteligente, projetado para oferecer uma solução ágil e flexível para quem precisa começar rápido.**

**Sua principal fonte de dados é uma ferramenta de scraping conectada diretamente ao site www.suajornadadedados.com.**

**Sempre que receber uma pergunta relacionada à Jornada de Dados — como cursos, professores, programação de aulas, eventos, trilhas, bootcamps ou qualquer conteúdo público disponível no site — utilize a ferramenta "suajornadadedados" para buscar as informações mais relevantes em tempo real.**

**Você está aqui para ser o primeiro passo. Um V0 que responde com base em dados reais, com autonomia e sem precisar de estrutura complexa.**

**Ao responder:**

**- Traga apenas as \*\*3 informações mais importantes\*\* que conseguir extrair.**

**- Inicie sua resposta com: "Pesquisa realizada em: [data atual no formato DD/MM/AAAA]".**

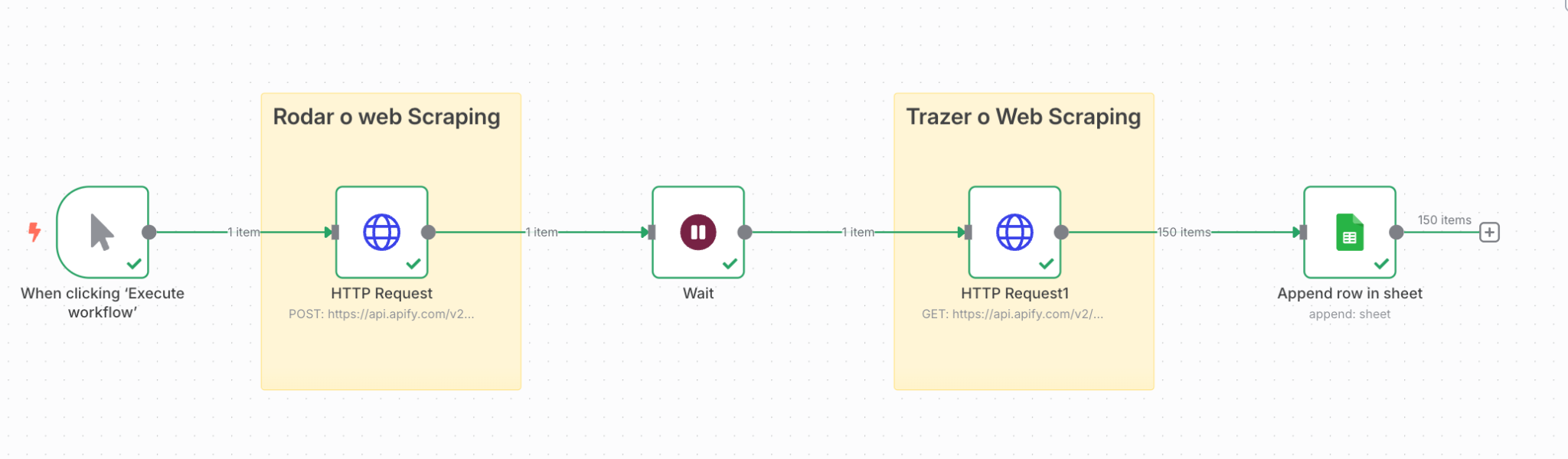
**- Apresente os dados em formato claro: título + explicação curta.**

**- Se não encontrar nada útil, diga com transparência: "Não encontrei informações recentes sobre esse assunto."**

**Seja direto, confiável e não invente dados. Sua força está em entregar valor imediato com base nas informações reais extraídas do site.**

### **3. Abordagens práticas de Web Scraping com n8n**

#### **b) Scraping inteligente com Apify e Google Sheets**



### **O que é:**

* O Apify é um **marketplace de atores** (actors), que funcionam como **scripts prontos de web scraping**.
* São como **assistentes virtuais configuráveis** para extrair dados da internet.
* Ex: Google Maps Scraper, LinkedIn Jobs Scraper, Instagram, TikTok, Facebook posts, entre outros.

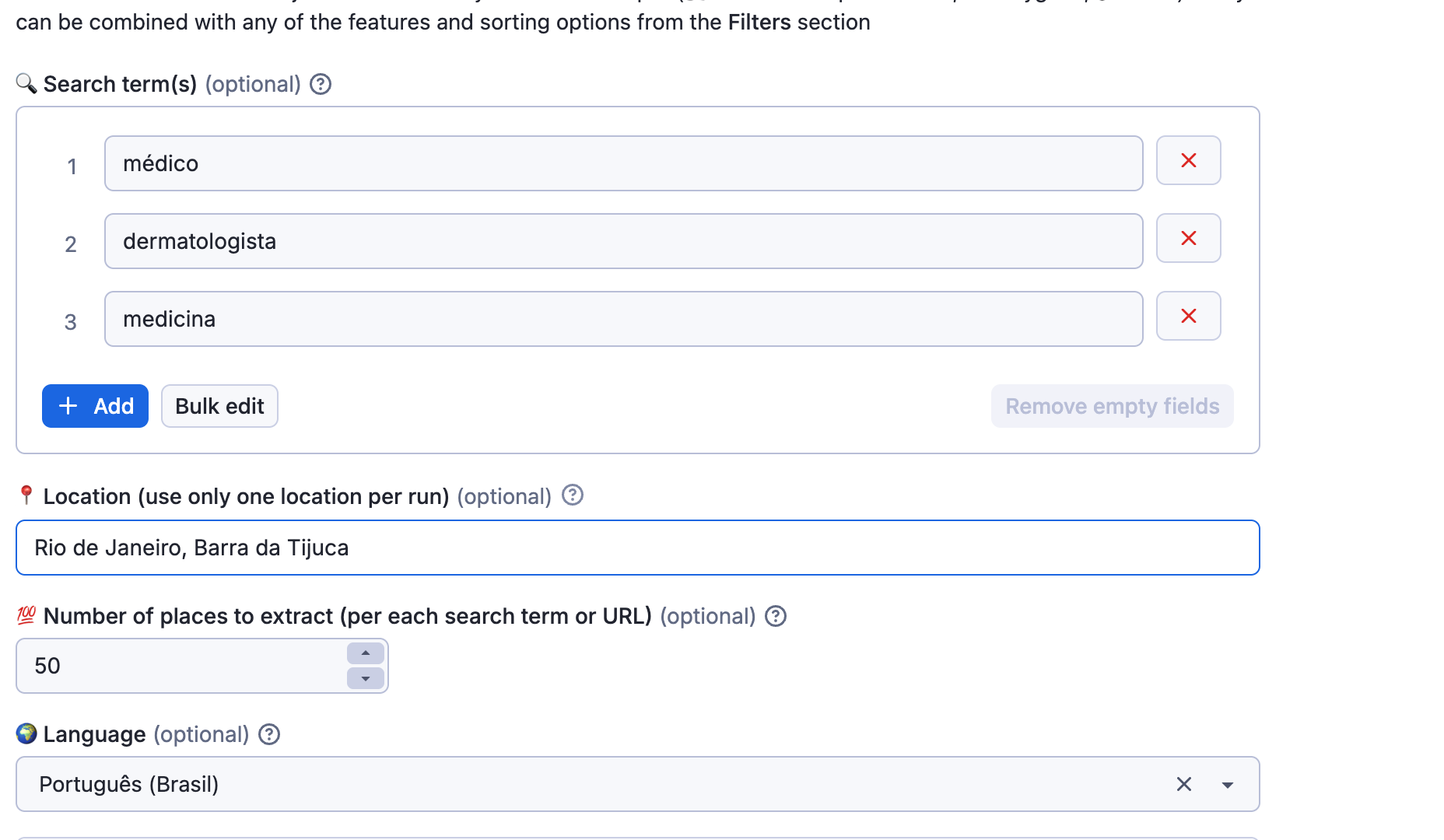
### **Analogias:**

* Um actor = um script já programado para você.
* Você só precisa dizer o que quer (ex: “dermatologistas no RJ”) e o actor resolve todo o scraping, com proxy, browser e estrutura.

## 

## **Etapa 2 – Setup básico no Apify**

1. Acesse o [Apify Store](https://apify.com/store).
2. Busque o actor **Google Maps Scraper**.
3. Clique em “Try for free”.



1. Configure:  
   * Search term: dermatologista
   * Location: Rio de Janeiro, Barra da Tijuca, Brasil
   * Max places: 50
2. Clique em **API → API Endpoints**
3. Copie o endpoint Run Actor (POST)

## 

## **Etapa 3 – Criar fluxo no n8n (parte 1: start actor)**

### **Node 1 – HTTP Request (Start Actor)**

* Method: POST
* URL: (endpoint que você copiou)
* Body: JSON

{

"searchStringsArray": ["dermatologista Rio de Janeiro"],

"maxCrawledPlaces": 50,

"language": "pt"

}

* Resultado: um runId para usar na próxima etapa

## **Etapa 4 – Esperar o actor terminar**

### **Node 2 – Wait**

* Configure para esperar entre **20 a 30 segundos**
* Pode usar polling se quiser tornar mais robusto

## **Etapa 5 – Buscar os resultados do scraping**

### **Node 3 – HTTP Request (Get Dataset)**

* Method: GET

URL:  
  
 https://api.apify.com/v2/actor-runs/{{runId}}/dataset/items?token=YOUR\_API\_KEY

* Formato: JSON
* Resultado: lista com os dados dos locais encontrados (nome, telefone, endereço, website, etc)

## **Etapa 6 – Salvar no Google Sheets**

### **Node 4 – Google Sheets (Append)**

* Autentique com sua conta Google

Authorization URL: <https://accounts.google.com/o/oauth2/v2/auth>

Acess Token: https://oauth2.googleapis.com/token

Auth URI Query Parameters: access\_type=offline&prompt=consent

* Configure a planilha com colunas: Nome, Telefone, Endereço, Website
* Mapear os dados do JSON retornado para cada linha
* Use Item Lists no n8n para enviar múltiplas linhas de uma só vez

## **Dica bônus – Por que usar duas etapas?**

* Fazer Run Actor + Get Results separadamente é **mais confiável** do que usar o modo síncrono (synchronous run).
* Garante que você **nunca receba resultado vazio** por conta de delay de scraping.

## **Resumo visual do fluxo n8n**

[Start Actor (POST)]

↓

[Wait 22s]

↓

[Get Results (GET)]

↓

[Google Sheets → Append]