

Challenge Sprint 1

**Coleta Automatizada de Dados de
Produtos HP em E-commerce (RPA)**

Conteúdo

1. Estratégia de Scraping	2
2. Fluxo de Navegação e Extração	2
2.1 Coleta de URLs de Produto (scrap_list)	2
2.2 Coleta de Dados do Produto (scrap_product)	2
2.3 Coleta de Comentários (scrap_comments)	3
3. Estrutura Geral do Código	3
4. Tabelas do Banco de Dados	4
4.1 products_ur	4
5. Desafios e Soluções	4
6. Evidências de Execução Bem-sucedida	5
7. Conclusão	6

Informações do Documento

Projeto:	Challenge Sprint 2025.1 – RPA – 1º semestre		
Equipe:	555677 - Matheus Hungaro Fidelis 556389 - Pablo Menezes Barreto 556984 - Tiago Toshio Kumagai Gibo 554668 - Israel Dalcin 555183 - Danilo Ramalho Silva 556213 - João Vitor Pires da Silva	Versão do Documento:	1.0

Histórico de versão

Versão	Data	Revisado por	Descrição	Nome do Arquivo
1.0	15/06	Tiago Toshio Kumagai Gibo	Criação do documento	Challenge Sprint 2025.1 - RPA

1. Estratégia de Scraping

- **Site-alvo:**
Mercado Livre (<https://www.mercadolivre.com.br>), principal marketplace do Brasil, selecionado por seu grande volume de produtos e avaliações, e por disponibilizar estrutura HTML relativamente estável.
- **Ferramenta:**
Python 3.8+ com Selenium WebDriver (ChromeDriver) para simular um navegador real, lidar com conteúdo dinâmico e interagir com elementos via JavaScript .
 - Uso opcional de modo headless (comentado em `main.py`).
 - Para evitar bloqueios e captchas, aplica-se variação de tempo de espera (`time.sleep` e `random.uniform`) e limites de páginas/produtos.

2. Fluxo de Navegação e Extração

2.1 Coleta de URLs de Produto (`scrap_list`)

- **Função:** `scrap_list(produto, url, driver)` em [mercadolivre.py](#).
- **Passos principais:**
 1. `driver.get(url)` e `time.sleep(2)` para aguardar carregamento e reduzir risco de bloqueio.
 2. Aceite de banner de cookies clicando em botão com classe `cookie-consent-banner-opt-out__action--key-accept`.
 3. Localização de `<ol class="ui-search-layout">` e, dentro dele, coleta de todos os títulos de produto em `h3.poly-component__title-wrapper` para extrair a URL via `a.get_attribute("href")`.
 4. Gravação de cada URL na tabela `products_url` com `save_url(produto, url)` .
 5. Paginação: encontra botão “Próximo” em `.andes-pagination__button--next a`, scroll até ele e dispara JavaScript para clicar após 3 s, retornando nova URL de página ou `False` se não houver próxima página.

2.2 Coleta de Dados do Produto (`scrap_product`)

- **Função:** `scrap_product(url, driver)` em [mercadolivre.py](#).
- **Dados extraídos:**
 - **Título e preço:** meta tag `og:title` via XPath, separando em "Título – Preço", convertendo

- preço para `float`.
 - **Avaliação média e quantidade de avaliações:** elementos em `div.ui-review-capability__rating`.
Descrição: texto de `p.ui-pdp-description__content`.
 - **Especificações técnicas:** itera linhas de `<tr>` em `div.ui-pdp-container__row--technical-specifications`, mapeando “marca”, “linha”, “modelo”, “formato de venda”, “conteúdo total em volume”, “rendimento de páginas” e “tipo de cartucho”.
 - **Vendedor:** texto em `div.ui-seller-data-header__title-container`.
- **Pós-processamento:** marca URL como `scraped = 1` na tabela `products_url`.
 - **Persistência:** insere registro em `products_data` via `save_product(...)`.

2.3 Coleta de Comentários (`scrap_comments`)

- **Função:** `scrap_comments(id, url, limit, driver)` em [mercadolivre.py](#).
- **Fluxo:**
 1. Acessa a página do produto e aguarda carregamento.
 2. Se existir botão “Mostrar mais” (`button.show-more-click`), clica e troca para `iframe#ui-pdp-iframe-reviews`.
 3. Laço até `limit` comentários ou fim: para cada comentário, extrai nota (`p.andes-visually-hidden`), data (`span.ui-review-capability-comments__comment__date`) e texto (`p.ui-review-capability-comments__comment__content`), salvando em `products_review` via `save_review(...)`.
 4. Atualiza `comments_scraped = 1` em `products_data`.

3. Estrutura Geral do Código

```

├── mercadolive.py    # Lógica de scraping (listas, produtos, comentários)
├── database.py       # Criação e manipulação de SQLite (tabelas e exportação)
├── main.py           # Fluxo de execução: inputs, inicialização do driver e chamadas às funções
├── requirements.txt  # Dependências (selenium, pandas, sqlite3 etc.)
├── readme.md         # Guia de instalação e uso :contentReference[oaicite:3]{index=3}
└── mercadolive.db    # Banco gerado após execução

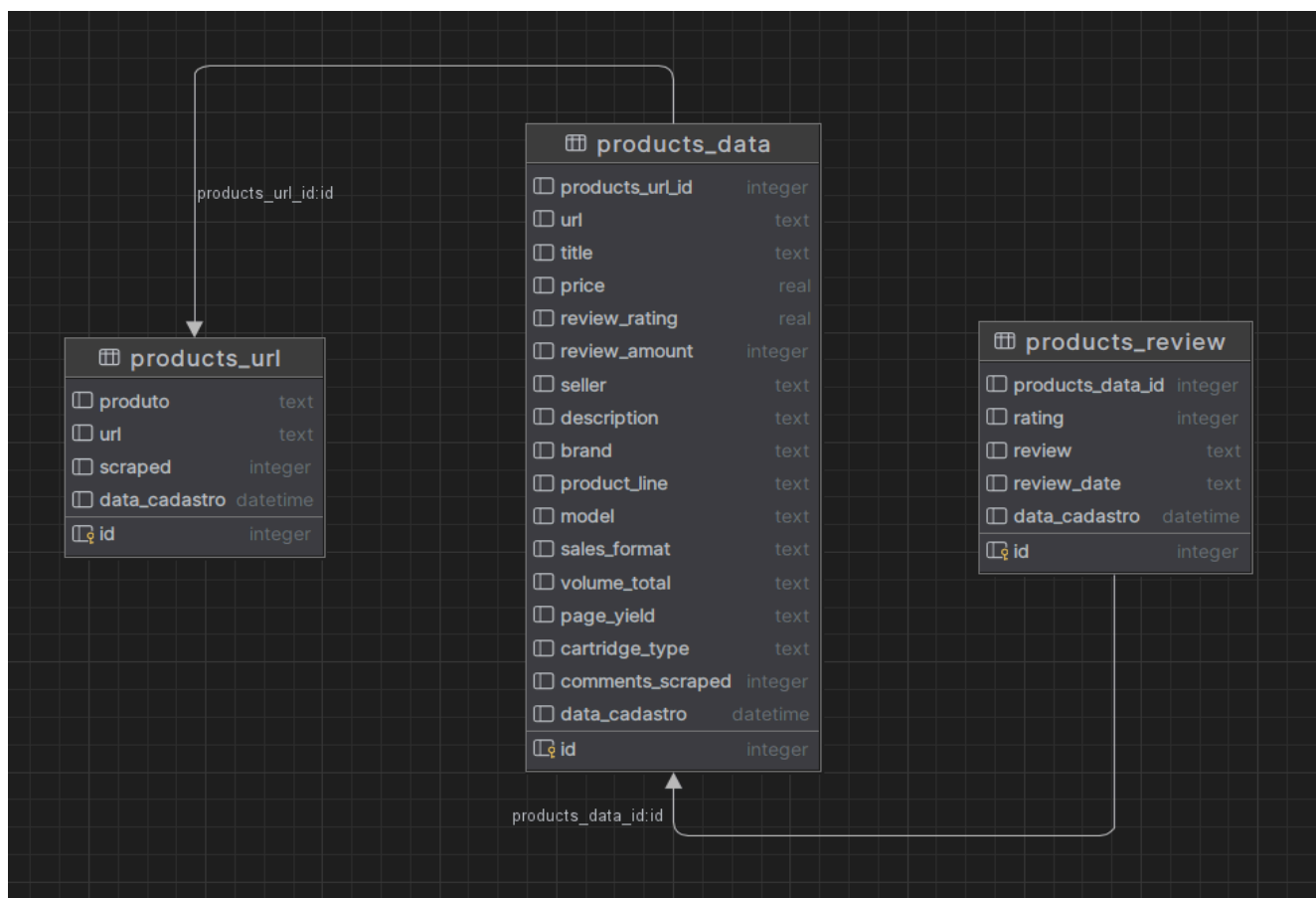
```

- **main.py:**

- Solicita ao usuário **produto**, **páginas**, **produtos** e **comentários**.
- Inicia ChromeDriver com opções configuráveis.
- Executa, na ordem: **list()**, **product()**, **comments()**, e por fim exporta CSVs em **outputs/** via **database.exportar_tabelas()** .
- **database.py:**
 - Tabelas **products_url**, **products_data**, **products_review**.
 - Funções **save_url**, **save_product**, **save_review**, **query**, **exportar_tabelas()** (gera CSVs).
- **requirements.txt:**
 - As versões utilizadas de **selenium**, **pandas**, **sqlite3** (via **python-dateutil**, **urllib3**, etc.) .

4. Tabelas do Banco de Dados

4.1 **products_ur**



5. Desafios e Soluções

Desafio	Estratégia de Solução
Paginação dinâmica	Uso de <code>time.sleep()</code> + <code>WebDriverWait</code> para garantir que o botão “Próximo” esteja visível antes de scrollar e clicar via <code>execute_script</code>
Conteúdo carregado por JavaScript	Combinação de <code>driver.implicitly_wait()</code> , <code>WebDriverWait</code> e <code>execute_script("scrollIntoView")</code> para forçar renderização de elementos antes da extração
Banner de cookies e pop-ups	Identificação e clique automático em botão de aceite (<code>cookie-consent-banner-opt-out__action--key-accept</code>)
Limite de requisições / bloqueios	Introdução de delays randômicos (<code>random.uniform(3.0, 4.0)</code>) entre acessos para simular comportamento humano.
Elementos ausentes / variações de layout	Tratamento via <code>try/except</code> em cada função de scraping, com logging de exceções e continuidade do fluxo sem interromper toda a execução.

6. Evidências de Execução Bem-sucedida

A seguir, trechos de registros extraídos diretamente do banco de dados (CSV gerados em `outputs/`):

Amostra de `products_url`

	id	produto	url	scraped	data_cadastro
1	1	cartucho hp 667 original	https://click1.mercadolivre.com.br/mclics/clicks/e...	1	2025-06-15 04:43:11
2	2	cartucho hp 667 original	https://click1.mercadolivre.com.br/mclics/clicks/e...	1	2025-06-15 04:43:11
3	3	cartucho hp 667 original	https://click1.mercadolivre.com.br/mclics/clicks/e...	1	2025-06-15 04:43:11
4	4	cartucho hp 667 original	https://www.mercadolivre.com.br/cartucho-hp-667-pr...	1	2025-06-15 04:43:11
5	5	cartucho hp 667 original	https://www.mercadolivre.com.br/cartucho-de-tinta-...	1	2025-06-15 04:43:11
6	6	cartucho hp 667 original	https://www.mercadolivre.com.br/cartucho-hp-667-co...	1	2025-06-15 04:43:11
7	7	cartucho hp 667 original	https://www.mercadolivre.com.br/cartucho-de-tinta-...	1	2025-06-15 04:43:11
8	8	cartucho hp 667 original	https://produto.mercadolivre.com.br/MLB-2685036387...	1	2025-06-15 04:43:11
9	9	cartucho hp 667 original	https://www.mercadolivre.com.br/cartucho-de-tinta-...	1	2025-06-15 04:43:12
10	10	cartucho hp 667 original	https://click1.mercadolivre.com.br/mclics/clicks/e...	1	2025-06-15 04:43:12

Amostra de `products_data`

	id	products_url_id	url	title	price	review_rating	review_amount	seller
1	1	1	https://www.mercadolivre.com.br/cartucho-hp-667xl-pr...	Cartucho Hp 667xl Preto E Color. 02 Preto E 01 Color...	298.5	4.6	80	Park Ecom
2	2	2	https://www.mercadolivre.com.br/hp-667xl-cartucho-de...	HP 667XL Cartucho de tinta Preta 3ym80al	162.63	4.5	82	MVCARTUCHOS4
3	3	3	https://www.mercadolivre.com.br/kit-cartucho-de-tint...	Kit Cartucho De Tinta Hp 667xl Preto + Colorido	270	4.6	80	YUKIINFORMATICA
4	4	4	https://www.mercadolivre.com.br/cartucho-hp-667-pre...	Cartucho HP 667 Preto 2376 2776 6476	66.9	4.7	1,892	OBERO INFORMATICA
5	5	5	https://www.mercadolivre.com.br/cartucho-de-tinta-hp...	Cartucho De Tinta Hp 667 Deskjet Ink Advantage Cor P...	153.1	4.8	1,953	OBERO INFORMATICA
6	6	6	https://www.mercadolivre.com.br/cartucho-hp-667-colo...	Cartucho Hp 667 Colorido 3ym78ab	71.9	4.7	1,892	vanvan
7	7	7	https://www.mercadolivre.com.br/cartucho-de-tinta-hp...	Cartucho De Tinta Hp 667 Cor Preto Do 2 ML	64.69	4.8	3,384	Eshop
8	8	8	https://produto.mercadolivre.com.br/MLB-2685036387-k...	Kit Cartucho Tinta Original Hp 667 Preto/colorido 23...	154.9	4.7	3,755	OBERO INFORMATICA
9	9	9	https://www.mercadolivre.com.br/cartucho-de-tinta-hp...	Cartucho De Tinta Hp Tricolor 667	68.55	4.7	1,201	Eshop
10	10	10	https://produto.mercadolivre.com.br/MLB-3208197871-k...	Kit Cartucho Hp 667 667xl Deskjet 2776 Preto E Color	199.9	4.4	337	Park Ecom

Amostra de products_review

	id	products_data_id	rating	review	review_date	data_cadastro
1	1	1	Avaliação 4 de 5	As tintas duram pouco, mas acredito que seja um problema da hp e ...	13 jun. 2025	2025-06-15 04:49:20
2	2	2	Avaliação 5 de 5	Muito bom,vou voltar a comprar de novo.	29 mai. 2025	2025-06-15 04:49:22
3	3	3	Avaliação 5 de 5	Já sou cliente faz tempo.	28 mai. 2025	2025-06-15 04:49:24
4	4	4	Avaliação 5 de 5	Excelente!.	16 mai. 2025	2025-06-15 04:49:26
5	5	5	Avaliação 5 de 5	Excelente produto.	28 abr. 2025	2025-06-15 04:49:28
6	6	6	Avaliação 5 de 5	Funcionando bem.	25 abr. 2025	2025-06-15 04:49:31
7	7	7	Avaliação 5 de 5	Comprei para guardar, mas verifiquei que todos são genuínos. Esto...	21 abr. 2025	2025-06-15 04:49:33
8	8	8	Avaliação 5 de 5	Bom material.	14 mar. 2025	2025-06-15 04:49:35
9	9	9	Avaliação 5 de 5	Excelente.	16 dez. 2024	2025-06-15 04:49:37
10	10	10	Avaliação 5 de 5	Bom.	19 nov. 2024	2025-06-15 04:49:39

7. Conclusão

Este processo de automação permite replicar de forma robusta a extração de listas, detalhes de produtos e avaliações no Mercado Livre, centralizando dados em um banco SQLite e exportando-os para CSV. A combinação de Selenium com waits explícitos, controle de delays e tratamento de exceções garante resiliência a bloqueios, dinamicidade de conteúdo e variações de layout.

Com esse relatório e o código disponível, outro desenvolvedor pode instalar as dependências, configurar o ChromeDriver e rodar `main.py` para reproduzir integralmente o pipeline de scraping.