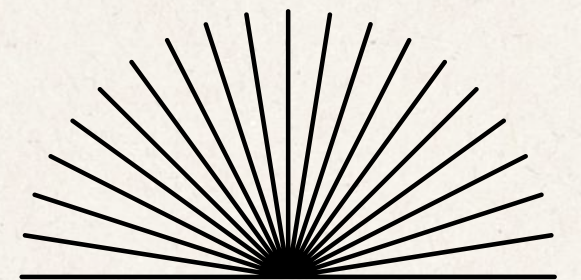




DATA PROJECT REPORT

Xu hướng thay đổi kì vọng công việc của các nhóm tuổi

PRESENTED BY:
Nguyễn Trang Anh



Agenda

I	Overview
II	Identify Data/Define Problem
III	Collect Data
IV	
V	
VI	
VII	



I. Overview

- **Title:** Tìm hiểu và phân tích về sự thay đổi kì vọng công việc của các ứng viên.
- **Scenario:** Khách hàng của công ty A đang gặp vấn đề với hiệu quả tuyển dụng. Công ty của khách hàng báo cáo rằng các bài đăng tuyển dụng của họ không thu hút đúng ứng viên và họ nhận được tỷ lệ không phù hợp cao khi phỏng vấn. Bộ phận nhân sự nghi ngờ các nhóm tuổi khác nhau có những kỳ vọng công việc hoàn toàn khác nhau, nhưng họ không có dữ liệu cụ thể về những gì mỗi nhóm tuổi thực sự tìm kiếm khi tìm việc.

II. Identify Data/Define Problem



1. Có thể truy cập dữ liệu nào trên thực tế?

- Nền tảng nào cho phép thu thập hoặc cung cấp API?
- Mức độ chi tiết trong dữ liệu tin tuyến dụng công khai?

2. Phương thức đảm bảo chất lượng và tính đầy đủ của dữ liệu?

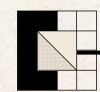
- Kích thước mẫu tối thiểu cho các kết quả có ý nghĩa thống kê?
- Xử lý dữ liệu bị thiếu hoặc định dạng không nhất quán
- Tính toán độ chính xác
- Các phương pháp xác thực đảm bảo độ chính xác?

3. Những ràng buộc về pháp lý và đạo đức là gì?

- Có giới hạn điều khoản dịch vụ đối với việc thu thập dữ liệu không?
- Bảo vệ quyền riêng tư cá nhân trong khi phân tích các mẫu?
- Những chính sách lưu giữ và chia sẻ dữ liệu?

Progress

Briefly discuss the key points for the project.



Identify



Collect Data

- Nguồn thu thập data
- Tool scrape data
- Lưu trữ data



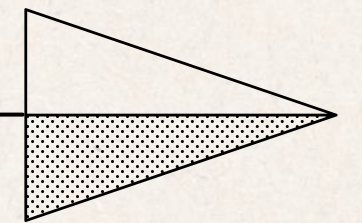
Clean Data



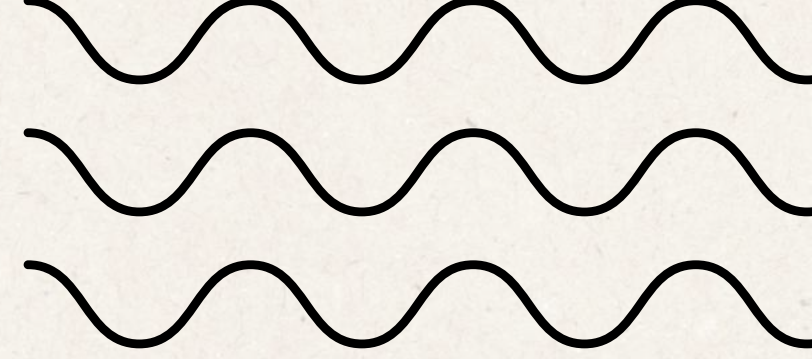
Analyze



Interpret



III. Collect Data



Một số khó khăn trong thu thập data

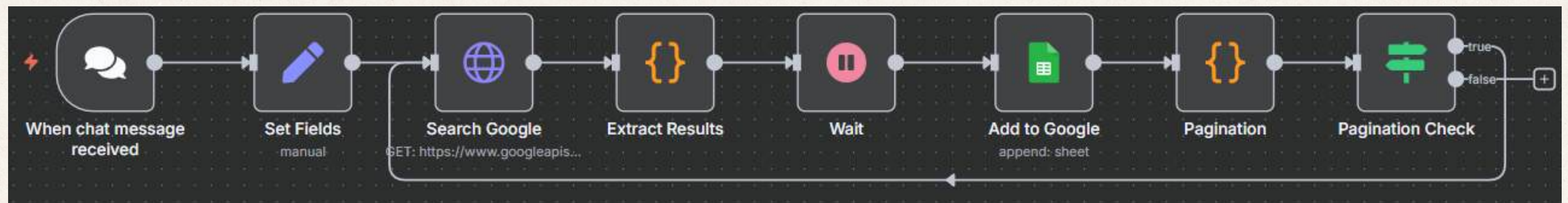
- Đa số các nền tảng tuyển dụng không public thông tin ứng viên, công ty chỉ nhìn thấy thông tin ứng viên khi họ apply vào công ty => Thiếu data real CV
- Một số ứng viên đăng CV file ảnh và PDF lên các nhóm tuyển dụng trên facebook => Cần nhắc thu thập CV trên các nhóm này và sử dụng OCR để extract thông tin
- Hiếm các dataset update thường xuyên về tình hình tuyển dụng của một nhóm cụ thể => có thể xảy ra thiếu mẫu

III. Collect Data

A. Scrape LinkedIn Profile từ kết quả Google Search

a. Tổng quan

- Xây dựng workflow trong n8n để extract kết quả google search LinkedIn profile và lưu vào trong Google Sheet
- Sử dụng add ons trong Google Sheet và link profile scrape được để scrape thêm thông tin chi tiết



III. Collect Data

A. Scrape Linkedin Profile từ kết quả Google Search

b. Tools

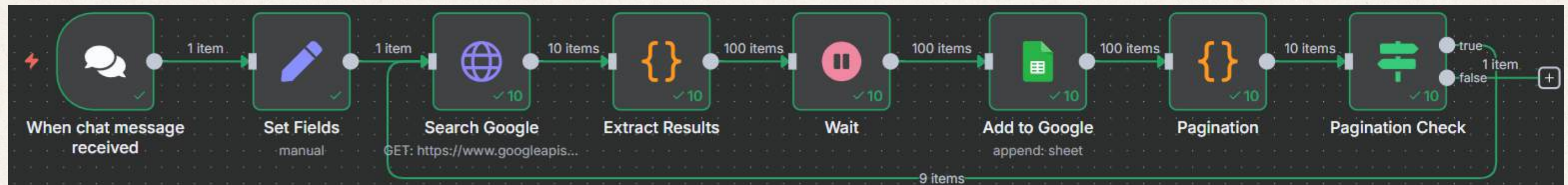
1.n8n

- **Định nghĩa:** n8n là một công cụ tự động hóa quy trình làm việc (workflow automation) mã nguồn mở, cho phép người dùng kết nối và tích hợp nhiều ứng dụng, dịch vụ khác nhau một cách trực quan mà không cần phải viết nhiều code.
- **Cơ chế hoạt động:** n8n vận hành dựa trên kiến trúc Node-based và Workflow-driven:
 - **Nodes:** Các khối chức năng độc lập, thực hiện một tác vụ cụ thể (ví dụ: gửi email, đọc dữ liệu từ database, gọi API).
 - **Workflows:** Chuỗi các Node được kết nối logic với nhau để tự động hóa một quy trình từ đầu đến cuối.
 - **Khả năng tích hợp:** tích hợp đa dạng ứng dụng và dịch vụ phổ biến (CRM, marketing, database, công cụ Dev, v.v.), cho phép xây dựng các quy trình tự động xuyên suốt hệ sinh thái ứng dụng

III. Collect Data

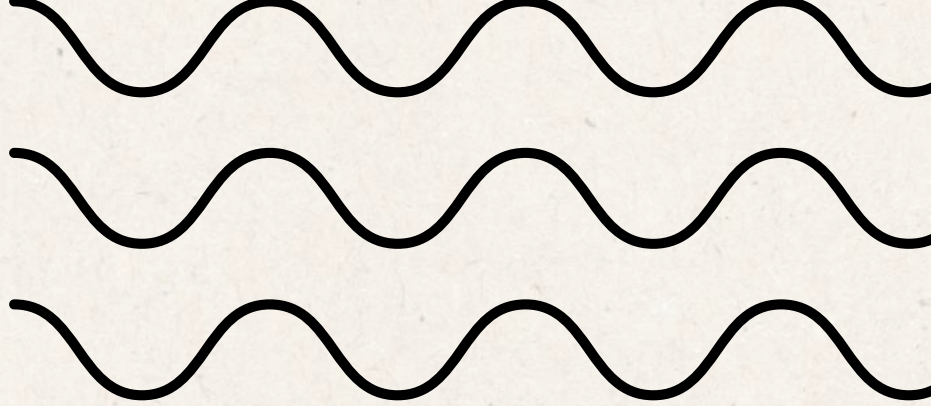
A. Scrape LinkedIn Profile từ kết quả Google Search

c. Giải thích workflow



III. Collect Data

A. Scrape Linkedin Profile từ kết quả Google Search

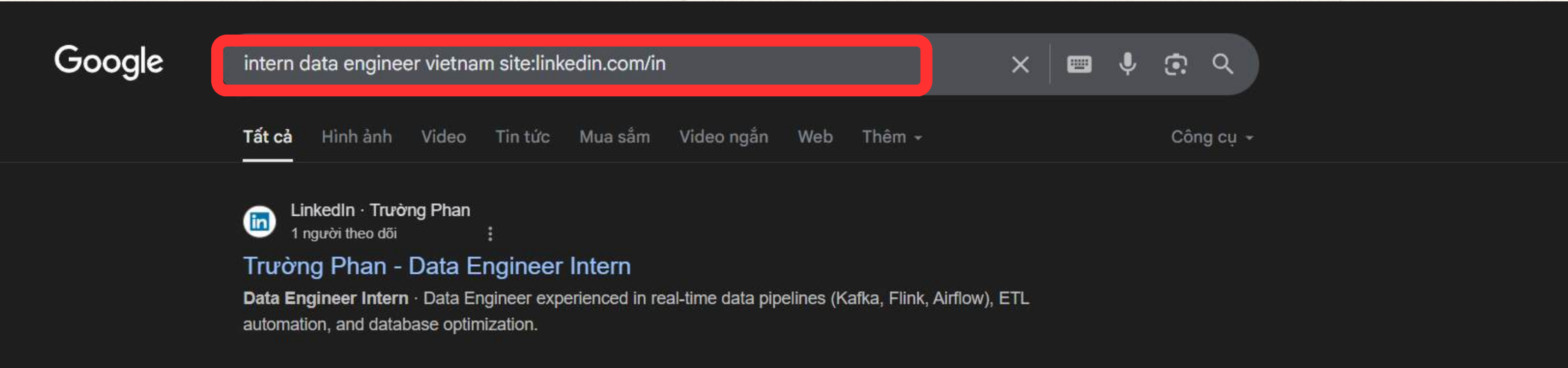


Node	Chức năng
When chat messages receive	Kích hoạt workflow khi tin nhắn được gửi đi
Set fields	<div>Xác định tham số<ul style="list-style-type: none">currentStartIndex: Vị trí bắt đầu trong kết quả tìm kiếmmaxPages: Số lượng trang lớn nhất khi có thể scrape trong 1 lần chạy workflow. Nếu maxPages = 10 và mỗi trang có 10 kết quả => Search 100 kết quả trong 1 lần chạy workflow</div>

III. Collect Data

A. Scrape Linkedin Profile từ kết quả Google Search

Node	Chức năng
Search Google	Thực hiện yêu cầu tìm kiếm của Google thông qua Google Custom Search API VD: Set value trong node Search Google là “intern data engineer vietnam site:linkedin.com/in”, workflow sẽ thực hiện search 10 trang đầu kết quả tìm kiếm (tham số maxPages của node Set Field) của Google



III. Collect Data

A. Scrape Linkedin Profile từ kết quả Google Search

Node	Chức năng
Extract Results	<p>Phân tích phản hồi API để chỉ trích xuất các trường liên quan. Chỉ giữ dữ liệu hữu ích và loại bỏ các chi tiết không cần thiết từ đầu ra API.</p> <p>Các trường dữ liệu</p> <ul style="list-style-type: none">• name: tên người dùng• title: tiêu đề của trang web hoặc tiêu đề kết quả tìm kiếm• link: đường liên kết trực tiếp đến trang• snippet: đoạn trích hoặc nội dung mô tả từ kết quả tìm kiếm• index: lấy số chỉ mục từ dữ liệu JSON nhằm theo dõi vị trí của kết quả tìm kiếm (phục vụ cho việc phân trang)• results: Kiểu Boolean (true/false) cho biết có nhiều kết quả tìm kiếm hơn hay không, cho phép workflow quyết định xem tiếp tục đến trang tiếp theo hay dừng.

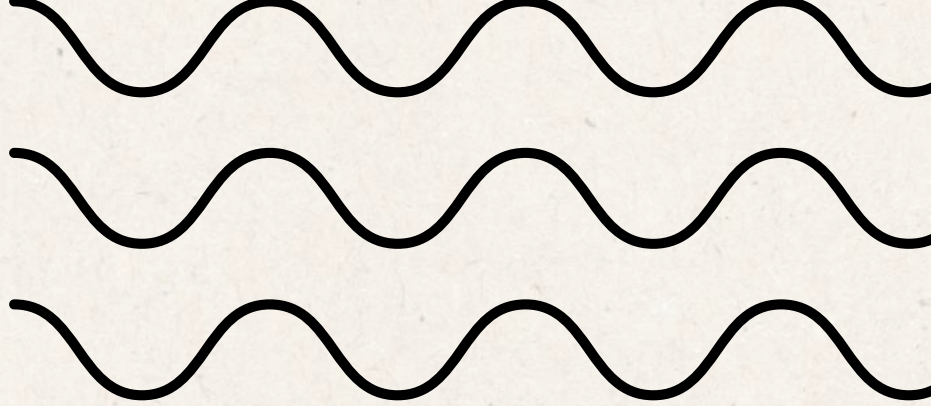
III. Collect Data

A. Scrape Linkedin Profile từ kết quả Google Search

Node	Chức năng
Wait	Tạm dừng một thời gian ngắn giữa các requests, tránh đạt giới hạn tốc độ API của Google và giảm nguy cơ bị quota bans
Add to Google (Sheet)	Gán kết quả tìm kiếm đã trích xuất vào Google Sheet
Pagination (Phân trang)	Cập nhật thông số tìm kiếm để tìm nạp nhóm kết quả tiếp theo, cho phép thu thập kết quả nhiều trang thay vì dừng lại ở trang đầu tiên.
Pagination Check	Quyết định xem có nên lặp lại và tìm kiếm trang tiếp theo hay dừng lại, kiểm soát khoảng cách phân trang trong workflow, tránh sử dụng quota không cần thiết

III. Collect Data

A. Scrape Linkedin Profile từ kết quả Google Search



A	B	C	D	E	F
name	title	link	snippet	index	results
Oliver Vu	Data Engineer Intern @ IDEXX O4U'25 Digital BAIS ...	https://www.linkedin.com/in/oliver-vu/	Data Engineer Intern @ IDEXX O4U'25 Digital BAIS @ USF F	11	TRUE
Duong (Rachel)	Junior Data Engineer @ Wedgewood LLC ...	https://www.linkedin.com/in/rachel-duong/	Experience ; Data Analyst Intern. Children's Hospital Los Angeles	11	TRUE
Minh Le	MA in Statistics at UC Berkeley Data Engineer Intern at ...	https://www.linkedin.com/in/minh-le/	Data Science Intern. AllxOne. Jun 2021 - Aug 2021 3 months. Ho	11	TRUE
Thanh Tran	SWE Data Science & Data Engineer Intern A.I ...	https://www.linkedin.com/in/thanh-tran/	SWE Data Science & Data Engineer Intern A.I Student Researc	11	TRUE

Ví dụ 4 mẫu dữ liệu scrape được và lưu trong Google Sheet khi chạy xong workflow

Link dữ liệu thu thập được:
[linkedin-profile](#)

III. Collect Data

A. Scrape LinkedIn Profile từ kết quả Google Search

c. Sử dụng addons trong Google Sheet mở rộng data

- Add ons LinkedIn Finder trong GG Sheet hỗ trợ scrape data chi tiết từ profile người dùng sử dụng link profile có trong sheet



LinkedIn, Email, Phon...

Gỡ cài đặt

Derrick™ is the all-in-one AI-powered extension for Google Sheets™ to import, enrich, and qualify B2B leads with emails, phone numbers, and LinkedIn™ data.

Nhà phát triển: [Derrick App](#) 

Đã cập nhật trang thông tin: 7 tháng 8, 2025

III. Collect Data

A. Scrape Linkedin Profile từ kết quả Google Search

c. Sử dụng addons trong Google Sheet mở rộng data

name	title	link	snippet	index	results	firstname (linked	lastname (linked	fullname (linked	location (linked	headline (linked	summary (linked	job title (linked
Duc Quang Tran	Front End Devel	https://www.linkedin.com/in/ducquangtran	Front End Devel	11	TRUE	Duc Quang	Tran	Duc Quang Tran	Gilbert, Arizona,	Front End Devel	Computer Scienc	Frontend Develc
Huy Anh Nguyen	Frontend Develc	https://ca.linkedin.com/in/huyanhnguyen	Since I started w	11	TRUE	Huy Anh	Nguyen	Huy Anh Nguyer	Burnaby, British	Frontend Develc	I am a Frontend	Frontend Develc
Sahar Mirkamran	Frontend Develc	https://www.linkedin.com/in/saharmirkamrani	Frontend Develc	11	TRUE	Sahar	Mirkamrani	Sahar Mirkamrani		Frontend Develc	I am a develop	Front-end devel

Kết quả sau khi sử dụng add ons

d. Tự đánh giá

- Giới hạn free API requests (100 requests/day), add ons giới hạn credit (200 free credits/month) => Chỉ scrape được 1000 profiles/ngày và chỉ mở rộng được 200 profiles
- Mỗi lần search ngành nghề khác nhau cần set value thủ công, chưa kết nối được AI Agent để tự động hóa quy trình scrape hằng ngày