

NHÓM 5

**HỆ THỐNG DỰ ĐOÁN ĐIỂM CHUẨN
CÁC TRƯỜNG ĐẠI HỌC VÀ GỢI Ý
NGUYỄN VỌNG CHO KÌ THI THPT**

Nội dung chính

- I. Lý do chọn đề tài
- II. Mục tiêu của các giai đoạn
- III. Nguồn dữ liệu
- IV. Quy trình thu thập dữ liệu
- V. Tiền xử lý và chuẩn hóa
- VI. Kết quả đạt được
- VII. Giới thiệu các bước tiếp theo

I. LÝ DO CHỌN ĐỀ TÀI



Dữ Liệu Phân Tán

Thông tin điểm thi và điểm chuẩn nằm rải rác ở nhiều nguồn.
=> gây khó khăn lớn cho việc tra cứu và so sánh tổng thể.



Quyết định chủ quan

Học sinh thường chọn nguyện vọng dựa trên cảm tính hoặc lời khuyên thiếu cơ sở thống kê tin cậy



Giải pháp dữ liệu

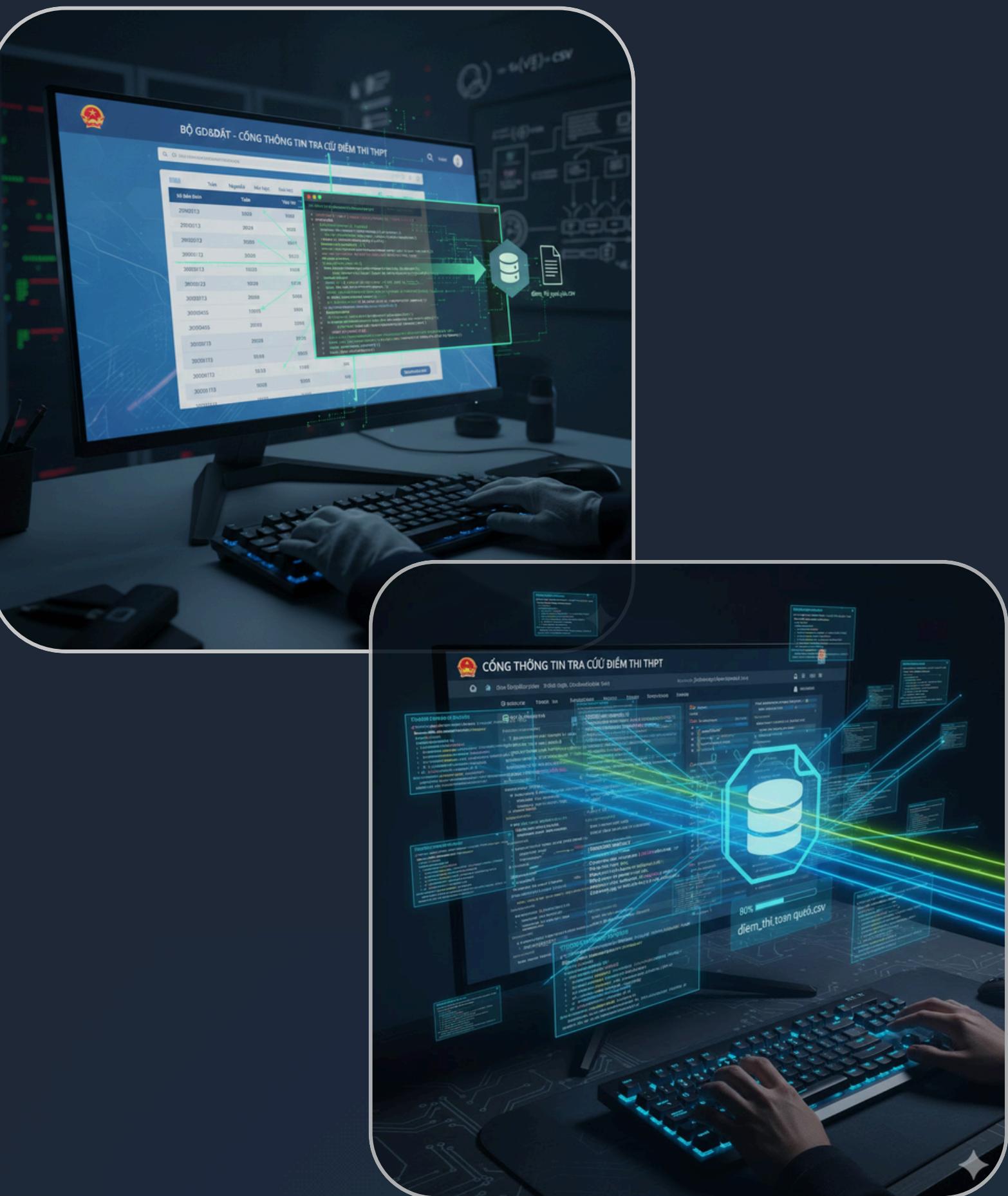
Xây dựng hệ thống tập trung, phân tích dữ liệu quá khứ để đưa ra gợi ý chính xác khách quan

II. MỤC TIÊU CỦA CÁC GIAI ĐOẠN



II.1 Thu Thập Dữ Liệu

- Thu thập điểm thi & điểm chuẩn từ các nguồn chính thống (Bộ GD&ĐT).
- Lưu trữ tập trung và chuẩn hóa dưới dạng CSV.
- Đảm bảo dữ liệu đầy đủ, đúng định dạng chi phân tích.



II. MỤC TIÊU CỦA CÁC GIAI ĐOẠN

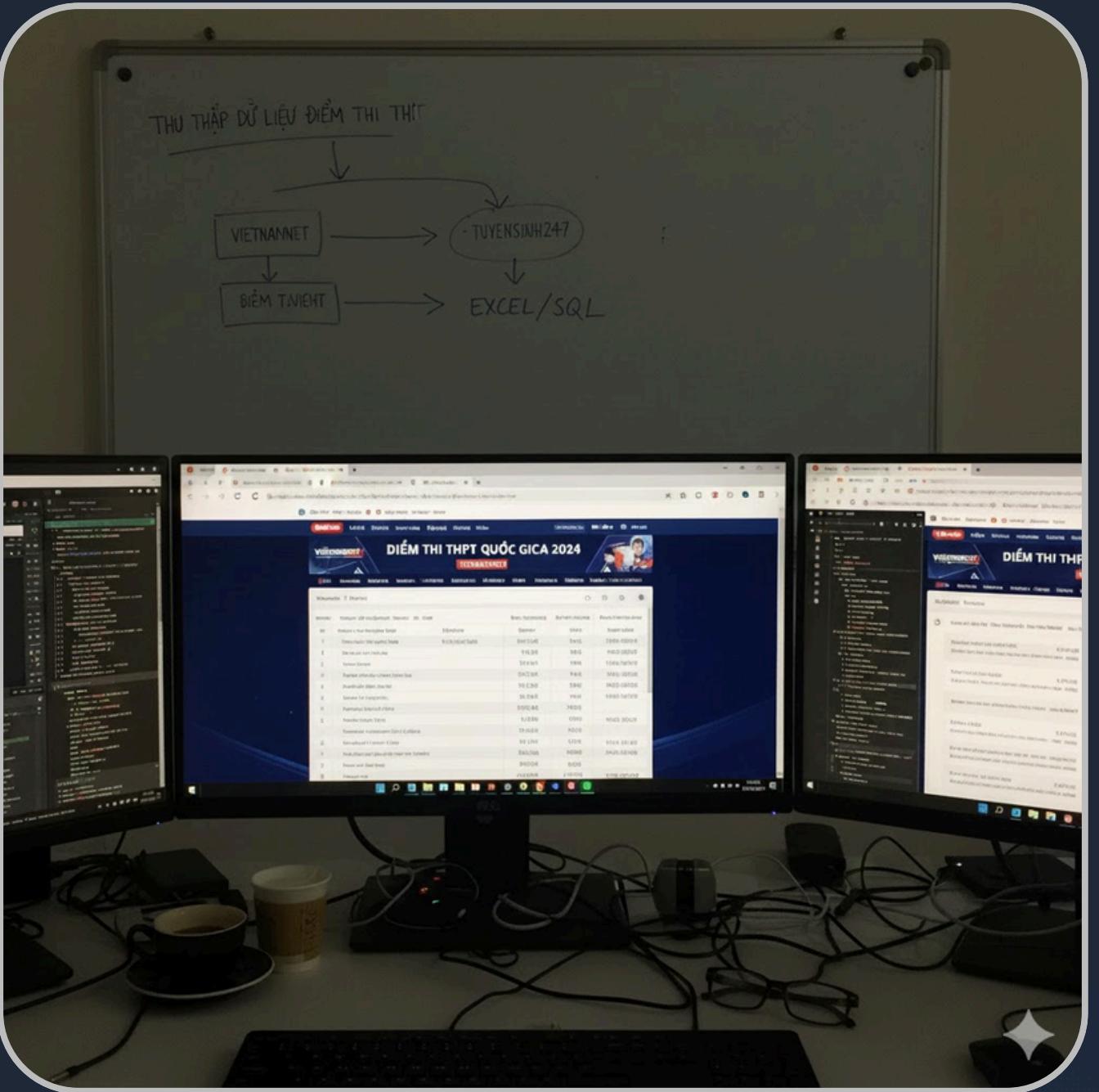
II.2 Tiết Xử Lý

-  **Làm sạch:** Loại bỏ trùng lặp, xử lý giá trị nhiễu/thiếu.
-  **Chuẩn hóa:** Đồng bộ mã trường, mã ngành toàn quốc.
-  **Tối ưu:** Tạo bộ dữ liệu chất lượng cao (Clean Dataset) cho huấn luyện mô hình.



III. NGUỒN DỮ LIỆU

-  VietnamNet: Tra cứu điểm thi tốt nghiệp THPT (2024).
-  Tuyensinh247: Tổng hợp điểm chuẩn đại học (2019-2024).



IV. QUY TRÌNH THU THẬP DỮ LIỆU

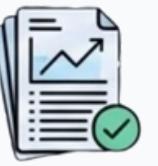


Danh Sách Trường

Crawl từ Tuyensinh247

```
GET /diem-chuan.html  
Parse: div.list-schol-box
```

Output: school.csv



Điểm thi THPT

Crawl từ Vietnamnet

```
Generate SBD (Mã tỉnh + ID)  
Multi-thread Requests
```

Output:
diem_thi_toan_quoc_
{nam_thi}.csv



Điểm chuẩn ĐH

Dựa trên danh sách các trường

```
Selenium Headless  
Handle "Xem thêm" button
```

Output: diem_chuan_all.csv

IV. QUY TRÌNH THU THẬP DỮ LIỆU



Tỉnh/Thành phố

Crawl từ Geocoding

Mapbox API

Rate limiting (sleep (1s), làm tròn 4 số thập phân).

Output: province.csv (Vĩ độ, Kinh độ).



Trường Đại Học

Crawl từ Geocoding

Mapbox API + Caching.

Dùng cache để tránh gọi lặp, sleep(1s) an toàn.

Output: school_with_coords.csv.

QUY TRÌNH CHI TIẾT

A. CRAWL DANH SÁCH TRƯỜNG ĐẠI HỌC

- Sử dụng **request** để gửi HTTP request và tải nội dung trang HTML.
- Dùng BeautifulSoup để phân tích HTML và tìm tất cả các thẻ `<div class="list-school-box">`, nơi chứa danh sách các trường
- Trích xuất thông tin trường.
- Lưu kết quả.
- Thông báo: sau khi hoàn tất, chương trình in ra số lượng trường đã lưu và đường dẫn file CSV.

QUY TRÌNH CHI TIẾT

A. CRAWL DANH SÁCH TRƯỜNG ĐẠI HỌC

Kết quả

- Dữ liệu được lưu vào file school.csv
- Thu thập được dữ liệu 300 trường cao đẳng đại học trên cả nước
- Dữ liệu gồm các feature: Mã trường, Tên trường, Link (dẫn đến nơi lấy điểm chuẩn).

QUY TRÌNH CHI TIẾT

B. CRAWL ĐIỂM THI THPT

- Mỗi thí sinh có SBD = Mã tỉnh (2 chữ số) + số báo danh 6 chữ số.
- Gửi request HTTP để tải trang kết quả từng SBD, sử dụng User-Agent giả lập trình duyệt.
- Dùng BeautifulSoup để phân tích HTML, tìm bảng điểm trong `<div class="resultSearch__right table">`.
- Trích xuất các thông tin điểm thi của thí sinh.
- Cơ chế crawl: sử dụng đa luồng (Multi-thread) quét song song theo batch 100 SBD.

QUY TRÌNH CHI TIẾT

B. CRAWL ĐIỂM THI THPT

Kết quả

- Dữ liệu được lưu vào file diem_thi_toan_quoc_{năm_thi}.csv
- Thu thập thành công dữ liệu bằng việc crawl kết hợp với các nguồn dữ liệu có sẵn.
- Dữ liệu gồm các feature: Năm thi, Mã tỉnh, SBD và điểm số của các môn thi.
- Số liệu dữ liệu thu thập được:

2019	2020	2021	2022	2023	2024
882594	870517	992295	995441	1022060	1061605

QUY TRÌNH CHI TIẾT

C. CRAWL ĐIỂM CHUẨN THPT CỦA CÁC TRƯỜNG ĐẠI HỌC

- + Dựa trên danh sách trường đã được crawl trước đó, lấy từ cột Link trong file school.csv
- + Cách thực hiện:
 - Mỗi trường đại học có 1 trang chi tiết chứa các bảng điểm chuẩn tuyển sinh theo từng năm.
 - Dùng Selenium (headless Chrome) để mở trang, tương tác với các nút “Xem thêm” nếu cần để tải đầy đủ dữ liệu cũ (từ 2019 đến 2024).
 - Phân tích HTML bằng BeautifulSoup để tìm các bảng điểm chuẩn có tiêu đề chứa “Điểm chuẩn” và “Điểm thi THPT”.

QUY TRÌNH CHI TIẾT

C. CRAWL ĐIỂM CHUẨN THPT CỦA CÁC TRƯỜNG ĐẠI HỌC

+ Trích xuất thông tin:

- Mã trường: từ cột trong DataFrame của trường
- Năm xét tuyển: từ tiêu đề bảng hoặc heading chứa năm
- Mã ngành: mã ngành tuyển sinh
- Tên ngành: tên ngành tương ứng
- Tổ hợp môn: tổ hợp xét tuyển
- Điểm chuẩn: điểm chuẩn cho từng ngành
- Ghi chú: nếu có

+ Cơ chế crawl:

- Tự động click "Xem thêm" để hiển thị dữ liệu các năm trước nếu chưa thấy năm 2019.
- Crawl theo từng trường, tuần tự hoặc theo slice của DataFrame để dễ quản lý tiến trình.

- Kiểm tra trùng lặp: bỏ qua các dòng đã lưu trước đó dựa trên khóa gồm Mã trường

+ Năm + Mã ngành + Tên ngành.

+ Lưu kết quả:

- Ghi từng batch vào file CSV, ví dụ: diem_chuan_thpt_all.csv.

+ Thông báo:

- In ra tiến trình crawl, các năm đã thấy, số dòng mới lưu vào CSV và cảnh báo nếu chưa thấy năm 2019 hoặc không tìm thấy bảng..

QUY TRÌNH CHI TIẾT

C. CRAWL ĐIỂM CHUẨN THPT CỦA CÁC TRƯỜNG ĐẠI HỌC

Kết quả

- Dữ liệu được lưu vào file diem_chuan_all.csv
- Thu thập được dữ liệu 29539 dòng điểm chuẩn của 300 cơ sở đại học, cao đẳng trên toàn quốc từ 2019 → 2024
- Dữ liệu gồm các feature: Mã ngành, Tên ngành, Tổ hợp môn, Điểm chuẩn, Ghi chú và Năm xét tuyển.

QUY TRÌNH CHI TIẾT

D. CRAWL TỌA ĐỘ TỈNH/THÀNH PHỐ

Cách thực hiện (Execution)

API: Sử dụng Mapbox Geocoding API để tra cứu tọa độ từ tên tỉnh thành.

Quy trình: Lặp qua danh sách 63 tỉnh đã định nghĩa, gọi API, làm tròn tọa độ đến 4 chữ số thập phân, và áp dụng độ trễ nhỏ (`time.sleep(1)`) để tránh bị chặn API.

Kết quả (Result)

Tên file đầu ra: province.csv

Cấu trúc dữ liệu: File CSV chứa Mã tỉnh, Tên tỉnh, Vĩ độ (VI_DO), và Kinh độ (KINH_DO).

QUY TRÌNH CHI TIẾT

E. CRAWL TỌA ĐỘ CÁC TRƯỜNG ĐẠI HỌC

Cách thực hiện (Execution)

API: Sử dụng Mapbox Geocoding API tương tự.

Quy trình: Đọc file CSV chứa danh sách trường, lặp qua cột tên trường (name_col). Sử dụng bộ nhớ đệm (cache) để tránh gọi API lặp lại cho cùng một tên trường, và sử dụng độ trễ cao hơn (time.sleep(1)) để đảm bảo tuân thủ giới hạn của API.

Kết quả (Result)

Cấu trúc dữ liệu: File CSV được bổ sung thêm hai cột mới là Kinh Độ và Vĩ Độ cho từng trường.

V. TIỀN XỬ LÍ & CHUẨN HÓA

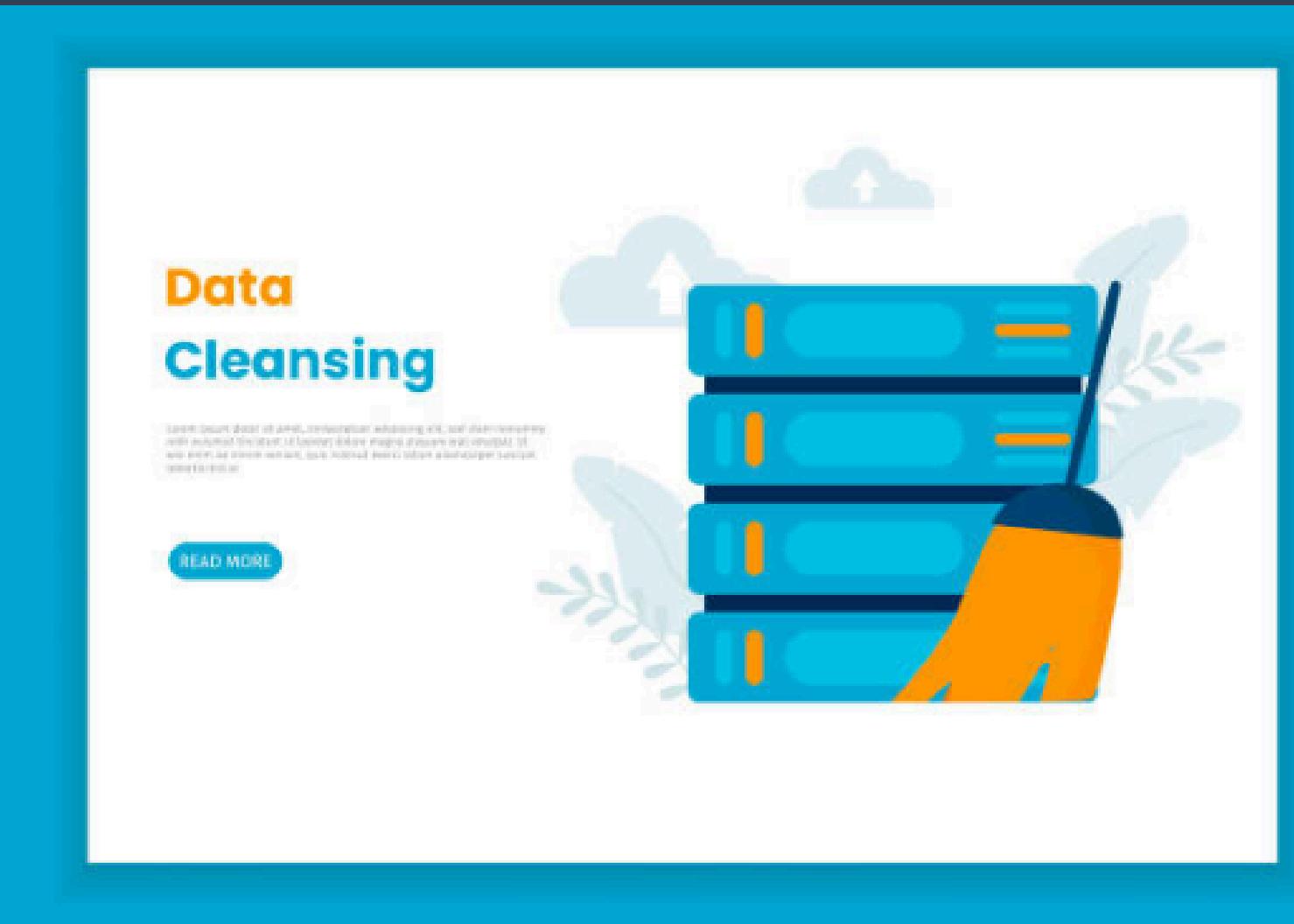
Làm sạch & Kiểm định dữ liệu

Xử lý thô

- ✓ **Chuẩn hóa định dạng SBD:** Đưa về format thống nhất (ví dụ: 8 chữ số).
- ✓ **Loại bỏ bản ghi trùng lặp:** Đảm bảo tính duy nhất cho mỗi thí sinh
- ✓ **Xử lý dữ liệu lỗi:** Loại bỏ hoặc điền khuyết các giá trị NaN/Null.

Lọc nghiệp vụ (Business Rules)

- ✓ **Loại bỏ điểm liệt:** Các điểm thành phần ≤ 1.0
- ✓ **Lọc điểm tổng thấp:** Loại bỏ các tổ hợp < 15 điểm để giảm nhiễu mô hình dự đoán.



Tối ưu dữ liệu điểm thi



Mã hóa không gian

Chuyển đổi tọa độ/địa chỉ thí sinh dưới dạng kinh độ/vĩ độ. Điều này hỗ trợ phân tích xu hướng điểm thi theo vùng miền địa lý một cách chính xác.



Chiến lược top 2

Tự động tính toán điểm cho mọi tổ hợp môn khả thi, nhưng chỉ lưu trữ và phân tích **2 khối thi cao nhất** của mỗi thí sinh để tập trung vào dữ liệu chất lượng.



Thống kê

Thay vì lưu chi tiết từng dòng điểm, hệ thống tổng hợp số lượng thí sinh theo phẩ điểm. Phương pháp này giúp **giảm tải dung lượng** lưu trữ đáng kể.

Chuẩn hóa dữ liệu điểm chuẩn

Đồng nhất chuỗi thời gian

Chỉ giữ lại các ngành học có dữ liệu điểm chuẩn đầy đủ và liên tục trong 6 năm (2019-2024) để đảm bảo độ chính xác khi training mô hình.

Chuẩn hóa và đồng bộ

Chuẩn hóa đồng bộ các **Mã ngành**.
Chuẩn hóa đồng bộ **Tên mã ngành**.

Mã ngành	Tên ngành	Tổ hợp môn
QHX01	Báo Chí	A00
QHX01	Báo Chí	C00
QHX01	Báo Chí	D01

Mã ngành	Tên ngành	Tổ hợp môn
QHX01_A00	Báo Chí	A00
QHX01_C00	Báo Chí	C00
QHX01_D01	Báo Chí	D01

Mã ngành	Tên ngành	Tổ hợp môn
7140233	Sư phạm	D01;D03;D96;D78
7140233	Sư phạm	D03
7140233	Sư phạm	D01;D78;D96

Mã ngành	Tên ngành	Tổ hợp môn
7140233	Sư phạm	D01;D03;D96;D78
7140233_D03	Sư phạm	D03
7140233_1	Sư phạm	D01;D78;D96

Mã ngành	Tên ngành
7520207	Kỹ thuật điện tử viễn thông
7520207	Kỹ thuật điện tử - viễn thông
7520207	Kỹ thuật điện tử - viễn thông

Mã ngành	Tên ngành
7520207	Kỹ thuật điện tử - viễn thông
7520207	Kỹ thuật điện tử - viễn thông
7520207	Kỹ thuật điện tử - viễn thông

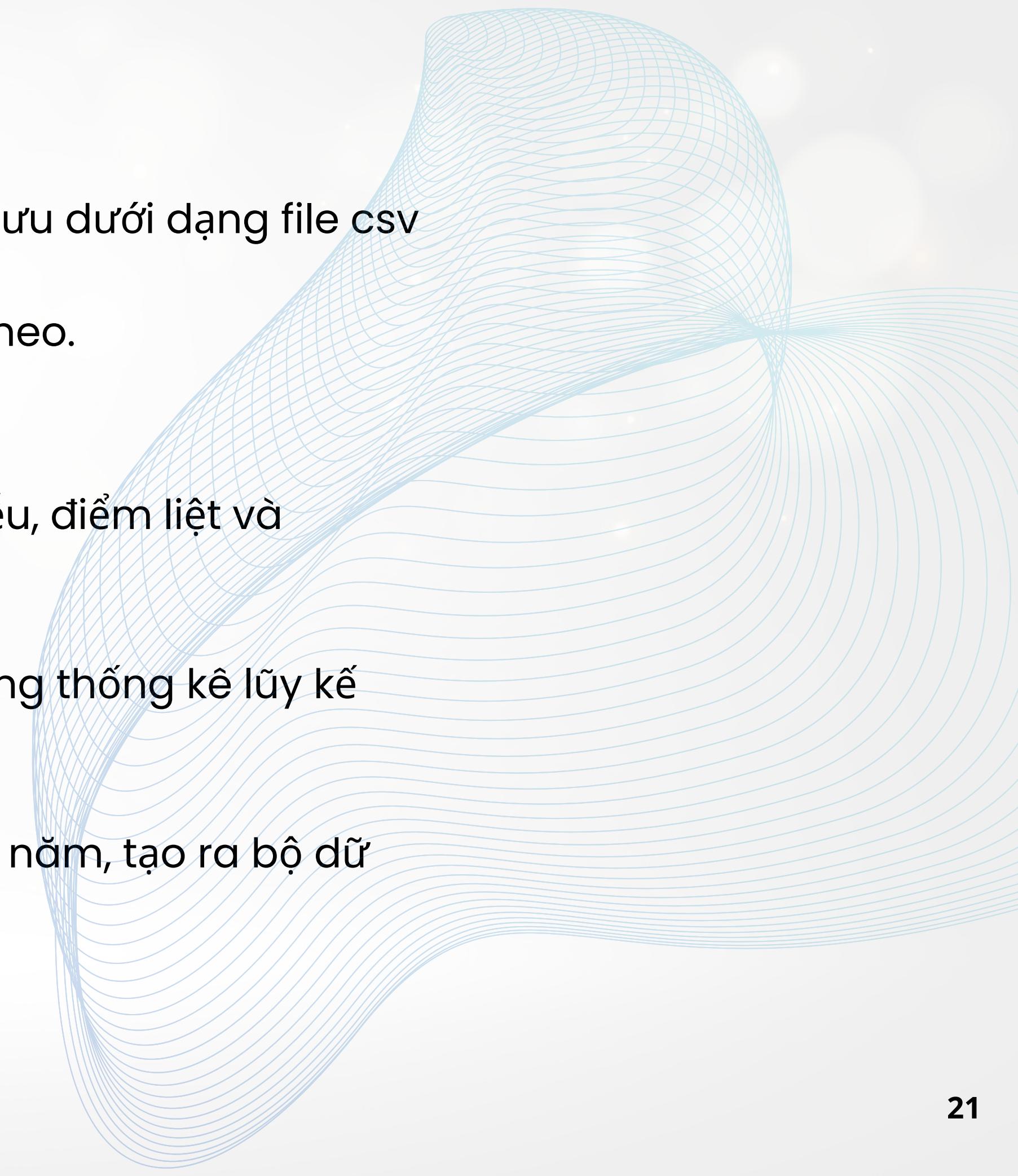
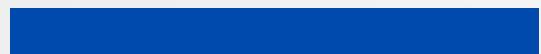
VI. KẾT QUẢ ĐẠT ĐƯỢC

- Thu thập dữ liệu:

- + Thu thập dữ liệu thành công, được chuẩn hóa và lưu dưới dạng file csv
- + Dữ liệu sẵn sàng để thực hiện các giai đoạn tiếp theo.

- Tiền xử lý dữ liệu:

- + **Làm sạch toàn diện:** Loại bỏ hoàn toàn dữ liệu nhiễu, điểm liệt và chuẩn hóa logic mã ngành.
- + **Tối ưu lưu trữ:** Chuyển đổi dữ liệu điểm thi sang dạng thống kê lũy kế giúp giảm tải dung lượng xử lý.
- + **Đồng bộ dữ liệu:** Đảm bảo chuỗi thời gian liên tục 6 năm, tạo ra bộ dữ liệu sạch sẵn sàng cho huấn luyện mô hình.



Tổng hợp số liệu sau khi lọc dữ liệu điểm thi

Năm	Số lượng ban đầu	Số lượng dữ liệu lỗi (SBD lặp/ rỗng)	Số lượng thí sinh bị điểm liệt (<= 1.0)	Số lượng thí sinh tổ hợp không đạt yêu cầu (< 15.0)	Số lượng sau khi lọc	Phần trăm dữ liệu đã loại bỏ
2019	882594	0	3671	108446	770477	12.70%
2020	882594	0	1276	40536	828705	4.80%
2021	992295	4591	1246	43259	943199	4.95%
2022	995441	0	1045	37585	956811	3.88%
2023	1022060	0	621	43712	977727	4.34%
2024	1061605	0	462	22512	1038631	2.16%

Một số kết luận từ thống kê trên

- Do ảnh hưởng của COVID-19 (thi đợt 2 + xét đặc cách) dẫn đến dữ liệu bị rỗng hoặc trùng lặp bất thường.
- Dữ liệu điểm thi năm 2024 là sạch nhất khi có tỉ lệ loại bỏ chỉ 2.16%.

TRÍCH CHỌN ĐẶC TRƯNG

Cột	Mô tả	Kiểu dữ liệu Kỹ Thuật/ Kiểu dữ liệu thống kê	Ý nghĩa
MA_TINH	Mã tỉnh thành	String/Nominal	Cho phép mô hình phân tích xu hướng cạnh tranh điểm số theo vùng miền.
NĂM_THI	Năm thi (2019-2024)	Integer/Discrete	Dùng để theo dõi sự thay đổi của phổ điểm qua các năm.
KHOI_THI	Các khối thi (A00, B00, D01...)	String/Nominal	Dữ liệu được phân tách theo từng khối xét tuyển độc lập.
MOC_DIEM	Mốc điểm (Step 0.05)	Float/Discrete	Đây là điểm chuẩn tiềm năng hoặc điểm ngưỡng để tính toán.
SO THI SINH	Phân vị thí sinh	Integer/Discrete	Đây là dữ liệu cho biết ở mốc điểm nhất định có bao nhiêu người đạt hoặc vượt qua.

LÝ DO LỰA CHỌN ĐẶC TRƯNG

MA_TINH: giúp phân tích và so sánh **phổ điểm theo khu vực địa lý** giữa các tỉnh thành

NAM THI: dùng để theo dõi xu hướng và sự thay đổi điểm qua các năm học

KHOI THI: phân tích độ cạnh tranh và nguồn lực của các lĩnh vực ngành nghề (Khoa học Tự nhiên, Xã hội, Kinh tế...).

MOC_DIEM: dùng làm ngưỡng cắt để đếm số lượng thí sinh

SO THI SINH: cho biết "Với mỗi mức điểm, có bao nhiêu người 'giỏi hơn' hoặc 'bằng' bạn". Đây là thông tin quyết định để đặt ngưỡng cắt (điểm chuẩn) cho quá trình tuyển sinh.

CHUẨN BỊ DỮ LIỆU CHO MÔ HÌNH

Cột	Mô tả	Kiểu dữ liệu Kỹ Thuật/ Kiểu dữ liệu thống kê
Mã trường	Mã viết tắt của trường đại học / học viện (ví dụ: QSX, HIU, ...)	String/Nominal (categorical)
Mã ngành	Mã số của ngành đào tạo theo quy định (ví dụ: 7380107, 7229030, ...)	String/Nominal
Tên ngành	Tên đầy đủ của ngành đào tạo (Luật kinh tế, Văn hóa học, ...)	String/Nominal
Tổ hợp môn năm 2019	Danh sách các tổ hợp xét tuyển năm 2019, dạng chuỗi “A00;D01;D14;...”	String/Nominal
Phân vị năm 2019	% thí sinh ≥ điểm chuẩn năm 2019.	Float/Continuous
...
Tổ hợp môn năm 2024	Danh sách các tổ hợp xét tuyển năm 2024, dạng chuỗi “A00;D01;D14;...”	String/Nominal
Phân vị năm 2024	% thí sinh ≥ điểm chuẩn năm 2024.	Float/Continuous

VII. GIỚI THIỆU CÁC BƯỚC TIẾP THEO.



Phân tích dữ liệu: thống kê, phân tích xu hướng và xác định ngành “hot”.



Hệ thống gợi ý (Recommendation System): dựa trên điểm dự kiến của thí sinh để đề xuất ngành/trường phù hợp, kèm xác suất trùng tuyển.

Trực quan hóa & báo cáo: biểu đồ, dashboard, chuẩn bị slide và demo hệ thống.



**THANK YOU
FOR LISTENING**