

**ĐẠI HỌC KHOA HỌC TỰ NHIÊN THÀNH PHỐ HỒ CHÍ MINH**

**CAO HỌC KHOA 30**

**NGÀNH KHOA HỌC DỮ LIỆU**



# MÔN KỸ THUẬT XỬ LÝ DỮ LIỆU

**GV: TS. NGUYỄN THANH BÌNH**

**Nhóm 11:**

Lê Chí Hoàng – 20C29021

Trần Mạnh Chánh Quân – 20C29014

Phạm Thị Hồng Phụng – 20C29033

## **Mục lục**

<b>Tổng quan</b>	<b>3</b>
1.1. Giới thiệu đề tài	3
1.2. Mục tiêu	3
<b>Phương pháp</b>	<b>4</b>
2.1 Tổng quan về data pipeline.	4
2.2 Thu thập dữ liệu.	4
2.3 Lưu dữ liệu vào database	7
2.4 Chạy Task Scheduler trên Windows để cập nhật dữ liệu mỗi ngày	11
2.5 Thiết kế dashboard trên Power BI	11
<b>Định hướng phát triển</b>	<b>13</b>

# 1. Tổng quan

## 1.1. Giới thiệu đề tài

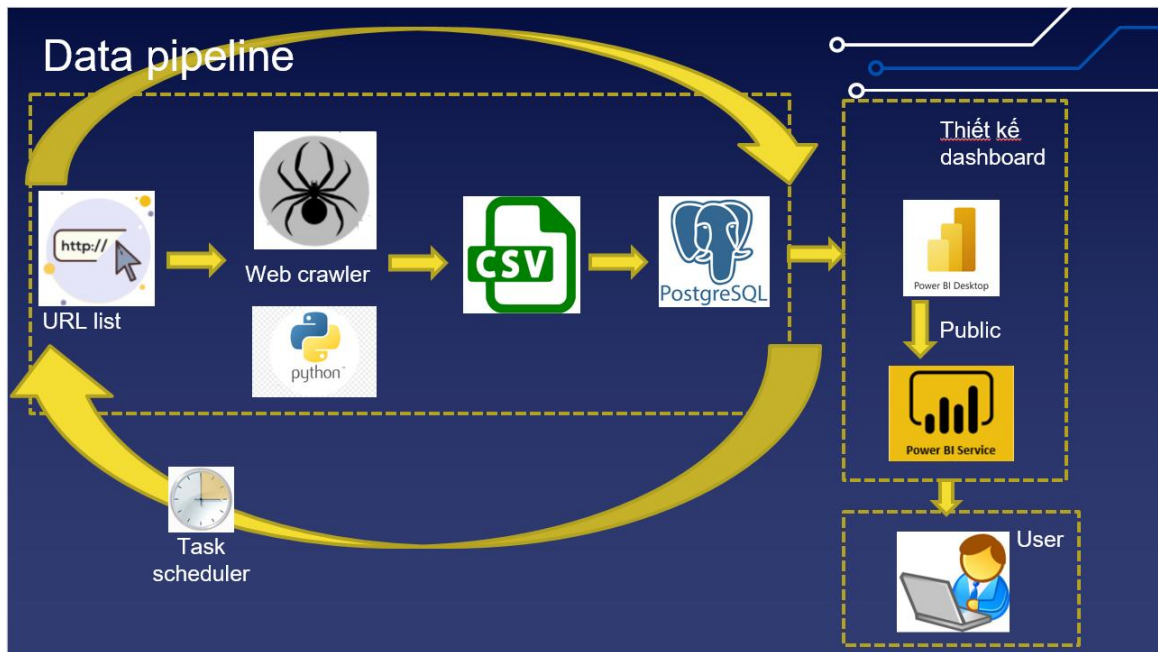
Thực trạng ô nhiễm môi trường không khí đang là vấn đề nhức nhối của thế giới và Việt Nam cũng không là ngoại lệ. Theo Báo cáo thường niên về chỉ số môi trường (The Environmental Performance Index - EPI) do tổ chức Môi trường Mỹ thực hiện, Việt Nam chúng ta là một trong 10 nước ô nhiễm môi trường không khí hàng đầu Châu Á. Tiêu biểu là ô nhiễm bụi (PM10, PM2.5). Thành phố Hồ Chí Minh là nơi bị ô nhiễm không khí nặng nhất của cả nước, có nhiều thời điểm bụi mịn (PM 2.5) bao phủ cả bầu trời làm hạn chế tầm nhìn, ảnh hưởng rất lớn đến sức khỏe của người dân. Do đó, việc theo dõi tình trạng ô nhiễm không khí là hết sức quan trọng và cấp bách, nhằm nâng cao nhận thức về ô nhiễm không khí, giúp mọi người hành động cải thiện chất lượng không khí và giảm thiểu tiếp xúc cá nhân với không khí ô nhiễm.

## 1.2. Mục tiêu

- Cung cấp những thông tin cụ thể về tình hình thời tiết, chất lượng không khí ở hai chỉ số AQI và PM2.5 để người dùng có cách phòng tránh phù hợp, bảo vệ sức khỏe.
- Góp phần nâng cao nhận thức của mọi người trong việc bảo vệ môi trường không khí.

## 2. Phương pháp

### 2.1 Tổng quan về data pipeline.

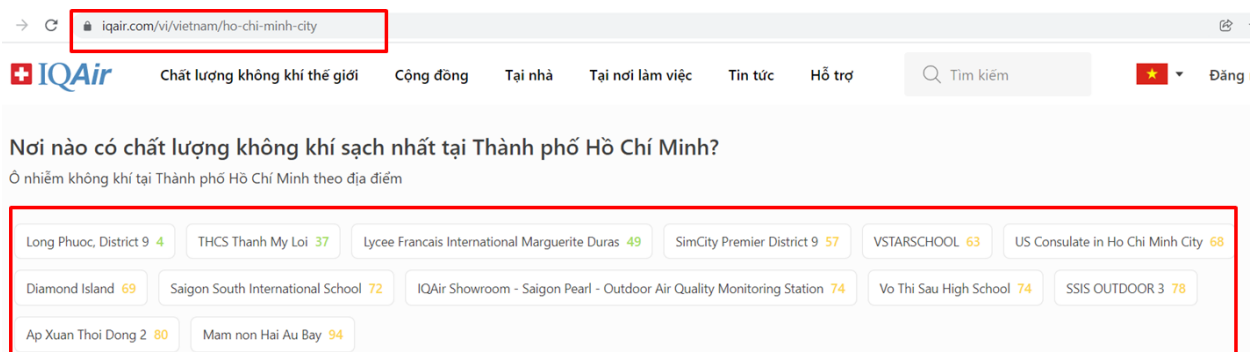


### 2.2 Thu thập dữ liệu.

#### 2.2.1 Công cụ thu thập: Selenium

Thông thường việc crawl data sẽ được thực hiện thông qua các API nhưng do website mà nhóm làm đề tài không hỗ trợ API nên nhóm đang crawl thông qua Selenium và code Python. Selenium là một thư viện và công cụ Python được sử dụng để tự động hóa việc thu thập dữ liệu từ các trình duyệt web. Một trong số đó là tính năng tìm kiếm trên web để trích xuất dữ liệu và thông tin hữu ích có thể không có sẵn.

Hiện tại nhóm không hardcode các URL của các trạm cố định, mà cách bot lấy data là đầu tiên bot vào website chính, tìm danh sách tất cả các trạm.



Sau đó bot sẽ vào từng trạm đọc data của các chart như bên dưới. Có bốn button tương ứng với bốn loại data: hàng giờ và hàng ngày theo chỉ số AQI, hàng giờ và hàng ngày theo chỉ số PM2.5.



### 2.2.2 Features

Data được thu thập từ website <https://www.iqair.com/vi/vietnam/ho-chi-minh-city> gồm một số thông tin chính như sau:

Attribute	Description
AQI	- Chỉ số chất lượng không khí





PM2.5	<ul style="list-style-type: none"> <li>- Chỉ số về chất lượng không khí, chỉ kích thước và mật độ những hạt trôi nổi trong không khí. Bụi PM2.5 là các hạt bụi lơ lửng có đường kính nhỏ hơn hoặc bằng 2,5 <math>\mu\text{m}</math> (micromet).</li> </ul>
location	<ul style="list-style-type: none"> <li>- Các trạm cung cấp dữ liệu chất lượng không khí tại TPHCM.</li> <li>- Hiện nay có tất cả 13 trạm: <ul style="list-style-type: none"> <li>- Long Phuoc, District 9</li> <li>- THCS Thanh My Loi</li> <li>- Lycee Francais International Marguerite Duras</li> <li>- SimCity Premier District 9</li> <li>- Diamond Island</li> <li>- IQAir Showroom - Saigon Pearl - Outdoor Air Quality Monitoring Station</li> <li>- Ap Xuan Thoi Dong 2</li> <li>- Vo Thi Sau High School</li> <li>- SSIS OUTDOOR 3</li> <li>- VSTAR SCHOOL</li> <li>- Mam non Hai Au Bay</li> <li>- US Consulate in Ho Chi Minh City</li> <li>- Saigon South International School</li> </ul> </li> </ul>
aqi_rating	<ul style="list-style-type: none"> <li>- Xếp loại theo chỉ số AQI, bao gồm 6 loại: <ul style="list-style-type: none"> <li>- Tốt (0 - 50)</li> <li>- Trung bình (51 - 100)</li> <li>- Không tốt đối với các nhóm nhạy cảm (101 - 150)</li> </ul> </li> </ul>

	<ul style="list-style-type: none"> <li>- Xấu (151 - 200)</li> <li>- Rất xấu (201 - 300)</li> <li>- Nguy hại (301 +)</li> </ul>
pm25_rating	<ul style="list-style-type: none"> <li>- Xếp loại theo chỉ số PM2.5, bao gồm 6 loại: <ul style="list-style-type: none"> <li>- Tốt (0 - 12,0)</li> <li>- Trung bình (12,1 - 35,4)</li> <li>- Không tốt đối với các nhóm nhạy cảm (35,5 - 55,4)</li> <li>- Xấu (55,5 - 150,4)</li> <li>- Rất xấu (150,5 - 250,4)</li> <li>- Nguy hại (250,5 +)</li> </ul> </li> </ul>

## 2.3 Lưu dữ liệu vào database

- Dữ liệu sau khi crawl về ta sẽ có được 13 thư mục tương ứng với 13 trạm, trong mỗi thư mục sẽ có 4 file dữ liệu, 2 file cho chỉ số chất lượng AQI (hourly và daily), 2 file cho chỉ số chất lượng PM2.5 (hourly và daily).

Name	Date modified	Type
Ap Xuan Thoi Dong 2	11/26/2021 4:03 PM	File folder
Diamond Island	11/26/2021 4:03 PM	File folder
IQAir Showroom - Saigon Pearl - Outdoo...	11/26/2021 4:03 PM	File folder
Long Phuoc, District 9	11/26/2021 4:03 PM	File folder
Lycee Francais International Marguerite D...	11/26/2021 4:03 PM	File folder
Mam non Hai Au Bay	11/26/2021 4:03 PM	File folder
Saigon South International School	11/26/2021 4:03 PM	File folder
SimCity Premier District 9	11/26/2021 4:03 PM	File folder
SSIS OUTDOOR 3	11/23/2021 12:24 PM	File folder
THCS Thanh My Loi	11/26/2021 4:03 PM	File folder
US Consulate in Ho Chi Minh City	11/26/2021 4:03 PM	File folder
Vo Thi Sau High School	11/26/2021 4:03 PM	File folder
VSTARSCHOOL	11/26/2021 4:03 PM	File folder

Name	Date modified
 aqidaily.csv	11/26/2021 9:28 PM
 aqihourly.csv	11/26/2021 9:29 PM
 pm2.5daily.csv	11/26/2021 9:28 PM
 pm2.5hourly.csv	11/26/2021 9:30 PM

- Tất cả các data này sẽ được xử lý tính toán/biến đổi để lưu vào 4 bảng trong database Postgres: stations, hourly\_measurement, daily\_report, ratings.
- Bảng **stations**:

Tên cột (Field Name)	Kiểu dữ liệu (Data Type)	Ghi chú
station_id(Khóa chính)	AutoNumber	Đây là ID của từng trạm
station_name_vn	varchar (256)	Tên trạm cung cấp dữ liệu chất lượng không khí.

- Bảng **hourly\_measurement**

Tên cột (Field Name)	Kiểu dữ liệu (Data Type)	Ghi chú
station_id(Khóa chính)	AutoNumber	ID của trạm tương ứng.
hourly_ts(Khóa chính)	timestamp	Thời gian tương ứng với chỉ số đo đạc thu được
pm25	real	Chỉ số chất lượng PM2.5
pm25_rating_id	AutoNumber	Đánh giá chất lượng là tốt hay xấu hay trung bình theo chỉ số PM2.5(tuy nhiên đây chỉ là ID, ta cần map qua bảng <b>ratings</b> để lấy được đánh giá



		này)
aqi	real	Chỉ số chất lượng AQI
aqi_rating_id	AutoNumber	Đánh giá chất lượng là tốt hay xấu hay trung bình theo chỉ số AQI(tuy nhiên đây chỉ là ID, ta cần map qua bảng <b>ratings</b> để lấy được đánh giá này)

- Bảng **daily\_report**

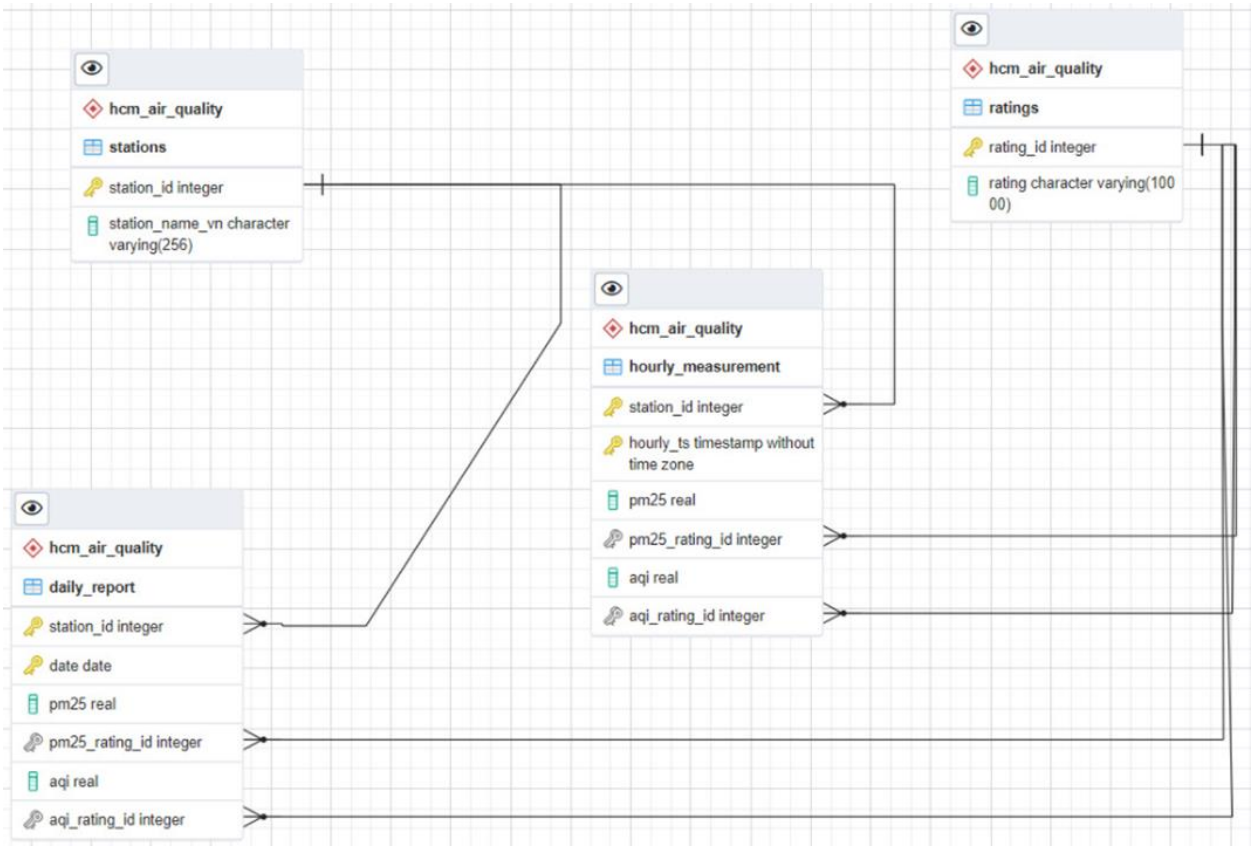
Tên cột (Field Name)	Kiểu dữ liệu (Data Type)	Ghi chú
station_id(Khóa chính)	AutoNumber	ID của trạm tương ứng.
date(Khóa chính)	date	Ngày tương ứng với chỉ số đo đạc thu được.
pm25	real	Chỉ số chất lượng PM2.5
pm25_rating_id	AutoNumber	Đánh giá chất lượng là tốt hay xấu hay trung bình theo chỉ số PM2.5(tuy nhiên đây chỉ là ID, ta cần map qua bảng <b>ratings</b> để lấy được đánh giá này)
aqi	real	Chỉ số chất lượng AQI
aqi_rating_id	AutoNumber	Đánh giá chất lượng là tốt hay xấu hay trung bình theo chỉ số AQI(tuy nhiên đây chỉ là ID,

		ta cần map qua bảng <b>ratings</b> để lấy được đánh giá này)
--	--	--

- Bảng **ratings**

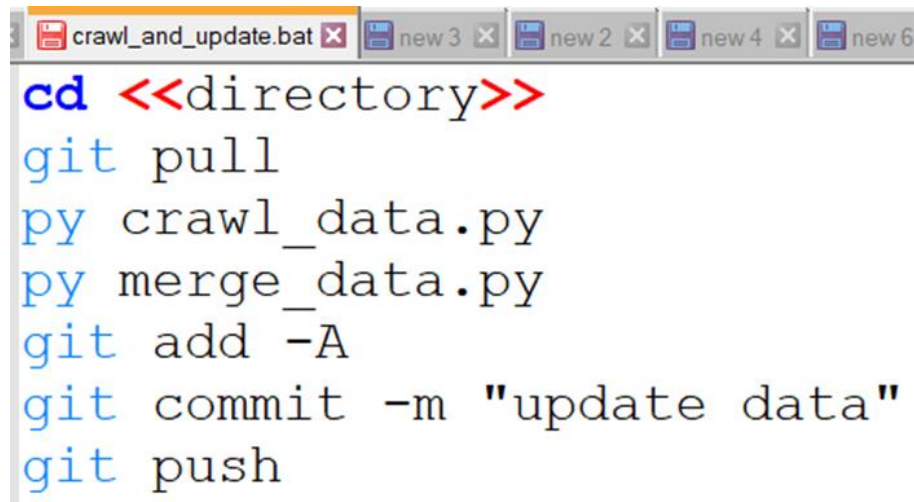
Tên cột (Field Name)	Kiểu dữ liệu (Data Type)	Ghi chú
rating_id(Khóa chính)	AutoNumber	ID của từng đánh giá, dùng để map với rating_id trong các bảng khác.
rating	varchar (10000)	Xếp loại chất lượng không khí tốt, xấu, trung bình, ...

- Mối liên kết giữa các bảng như sau:



## 2.4 Chạy Task Scheduler trên Windows để cập nhật dữ liệu mỗi ngày

- Task Scheduler cho phép thiết lập việc thực hiện một nhiệm vụ nào đó một cách tự động ở các mốc thời gian cài đặt trước đó bằng cách khởi động task và thực hiện nó khi các điều kiện cài đặt trước đó đúng. Ta sẽ dùng Task Scheduler để thực hiện việc crawl data và update data mỗi ngày.
- Tạo một file batch như bên dưới để tự động hoá việc cập nhật data. Sau đó khi thiết lập Task Scheduler thì sẽ dùng file batch này để thiết lập.

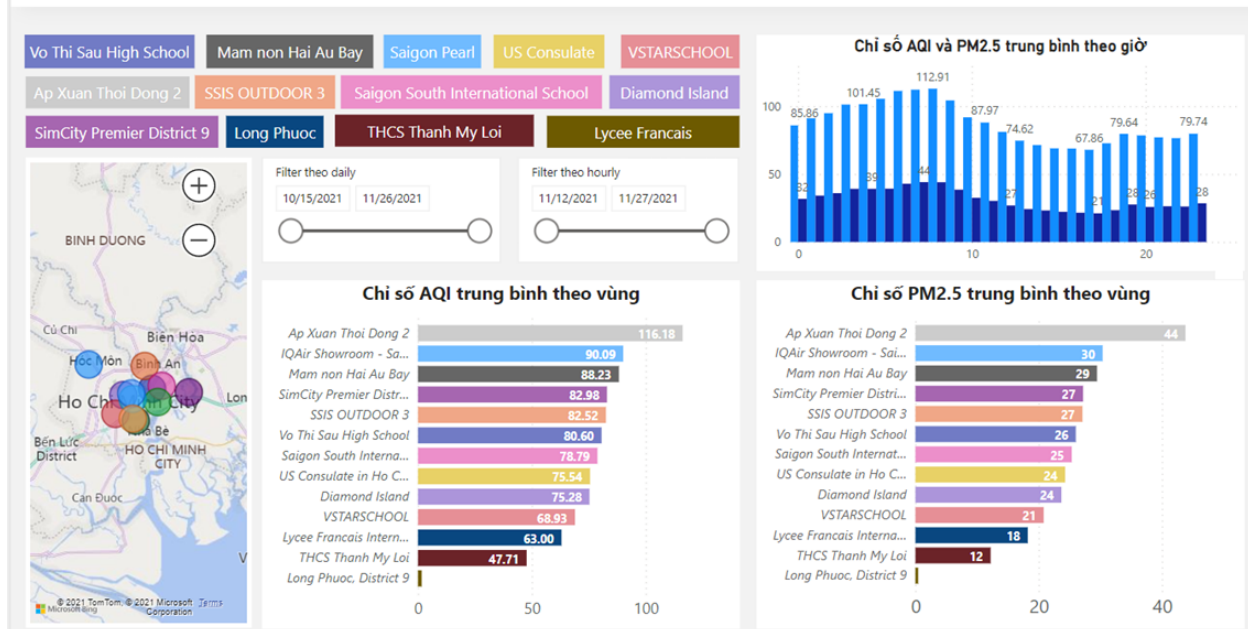


```
cd <<directory>>
git pull
py crawl_data.py
py merge_data.py
git add -A
git commit -m "update data"
git push
```

## 2.5 Thiết kế dashboard trên Power BI

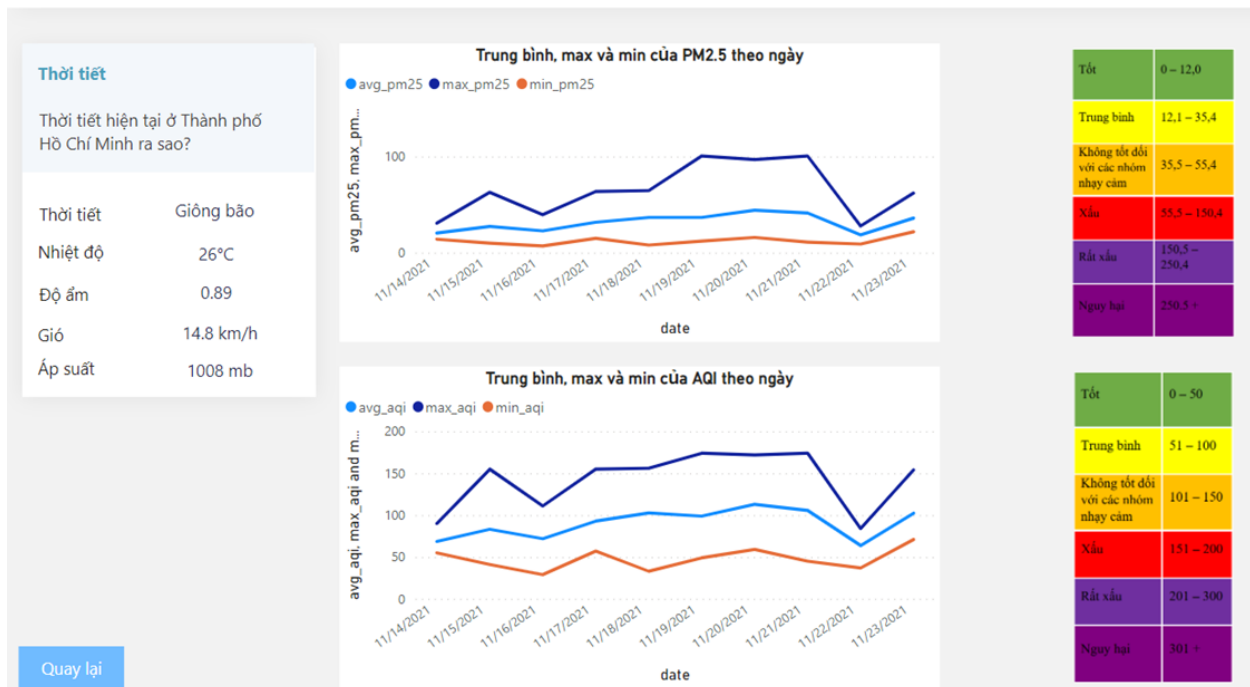
- Dashboard gồm hai trang là Home và Detail.
- Trang Home thể hiện các thông tin sau:
  - Chỉ số AQI và PM2.5 trung bình theo giờ
  - Chỉ số AQI trung bình cho tất cả các trạm
  - Chỉ số PM2.5 trung bình cho tất cả các trạm
  - Thời tiết hôm nay tại TPHCM

## Theo dõi tình trạng ô nhiễm không khí tại Thành phố HCM



- Trang Detail thể hiện các thông tin chi tiết của từng trạm như:
  - Hình 1: vẽ 3 đường cho AQI, đường có giá trị thấp nhất/trung bình/cao nhất của các ngày
  - Hình 2: chú thích 6 mức độ nguy hiểm của AQI
  - Hình 3: vẽ 3 đường cho PM2.5, đường có giá trị thấp nhất/trung bình/cao nhất của các ngày
  - Hình 4: chú thích 6 mức độ nguy hiểm của PM2.5

## Theo dõi tình trạng ô nhiễm không khí tại Thành phố HCM



### 3. Định hướng phát triển

- Phát triển thêm mô hình đưa ra các khuyến cáo, cảnh báo về chất lượng không khí cho các nhóm đối tượng có các bệnh như hen suyễn, viêm xoang, hô hấp,...
- Phát triển các mô hình để dự đoán mức độ của ô nhiễm không khí đối với sức khỏe cộng đồng bằng cách phân tích sự tương quan mức độ ô nhiễm không khí trong các khoảng thời gian ngắn hạn, trung hạn và dài hạn tại TP.HCM.
- Mở rộng quy mô ra khắp các tỉnh thành của Việt Nam.
- Hiện tại do data không đủ nên chỉ mới thống kê tình trạng ô nhiễm theo giờ, theo ngày, theo tuần, sau này sẽ tiếp tục thống kê theo tháng, theo quý, theo năm.
- Hiện tại do thời gian có hạn nên nhóm đang tạo ra database nhưng chưa áp dụng nhiều, chủ yếu khi xây dựng dashboard thì lấy data trực tiếp từ csv để thao tác. Trong tương lai nhóm sẽ nâng cấp thêm để dashboard sẽ lấy data trực tiếp từ database.