KHOA CÔNG NGHỆ THÔNG TIN
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
227 Nguyễn Văn Cừ, Phường 4, Quận 5, TP.HCM
Điện Thoại: (08) 38.354.266 - Fax:(08) 38.350.096

CSC17106 – XỬ LÝ PHÂN TÍCH DỮ LIỆU TRỰC TUYẾN

HƯỚNG DẪN THỰC HÀNH

AIR FLOW

I. Thông tin chung

Mã số:	HD03
Thời lượng dự kiến:	3 tiếng
Deadline nộp bài:	-
Hình thức:	-
Hình thức nộp bài:	-
GV phụ trách:	Phạm Minh Tú
Thông tin liên lạc với GV:	pmtu@fit.hcmus.edu.vn

II. Chuẩn đầu ra cần đạt

Bài hướng dẫn này nhằm mục tiêu đạt giúp sinh viên được các mục tiêu sau:

1. Cài đặt Air Flow trong môi trường window
2. Cấu hình Air Flow và tìm hiểu hệ thống quản trị
3. Tạo chương trình air flow dùng python.

III. Mô tả

Apache Airflow là một công cụ giúp bạn quản lý và lên lịch cho các đường ống dữ liệu.,nó cho phép "tự động viết mã, lên lịch và theo dõi quy trình làm việc."

Airflow là một công cụ quan trọng đối với các kỹ sư và nhà khoa học dữ liệu.
Mặc dù được khuyến nghị chạy Airflow với Docker, tuy nhiên phương pháp này hoạt động không tốt trên các máy có bộ nhớ thấp không thể chạy Docker.

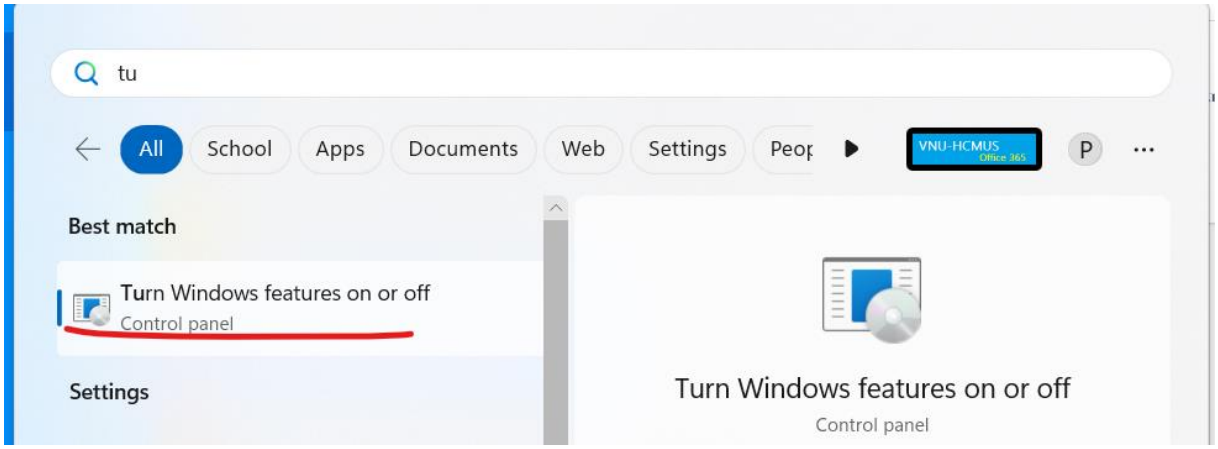
Yêu cầu:
Bạn cần có Python 3.8 hoặc cao hơn, Windows 10 hoặc cao hơn và Windows Subsystem for Linux (WSL2) làm theo bài hướng dẫn này,

Windows Subsystem for Linux (WSL2) là gì?
WSL2 cho phép bạn chạy các lệnh và chương trình Linux trên hệ điều hành Windows.

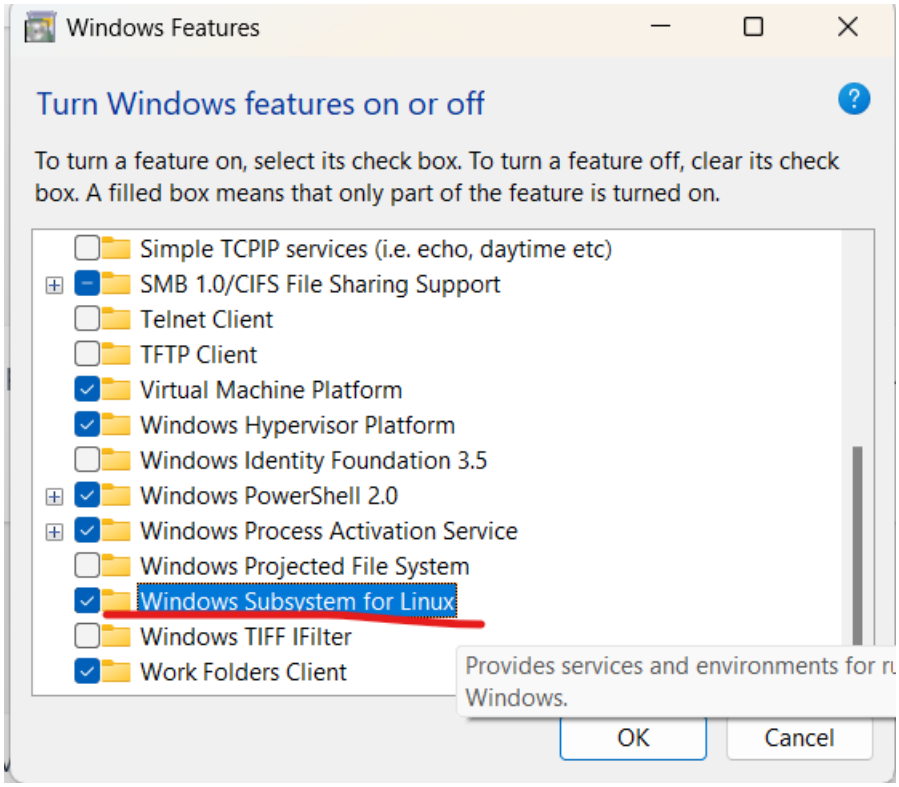
Nó cung cấp một môi trường tương thích với Linux chạy trực tiếp trên Windows, cho phép người dùng sử dụng các công cụ và tiện ích dòng lệnh Linux trên máy tính chạy Windows.

Hướng dẫn cài đặt WSL2:

Đầu tiên mở tính năng Windows Subsystem for Linux bằng tính năng **Turn Windows features on or off**



Chọn Windows Subsystem for Linux



Sau đó tiến hành cài đặt bình thường.

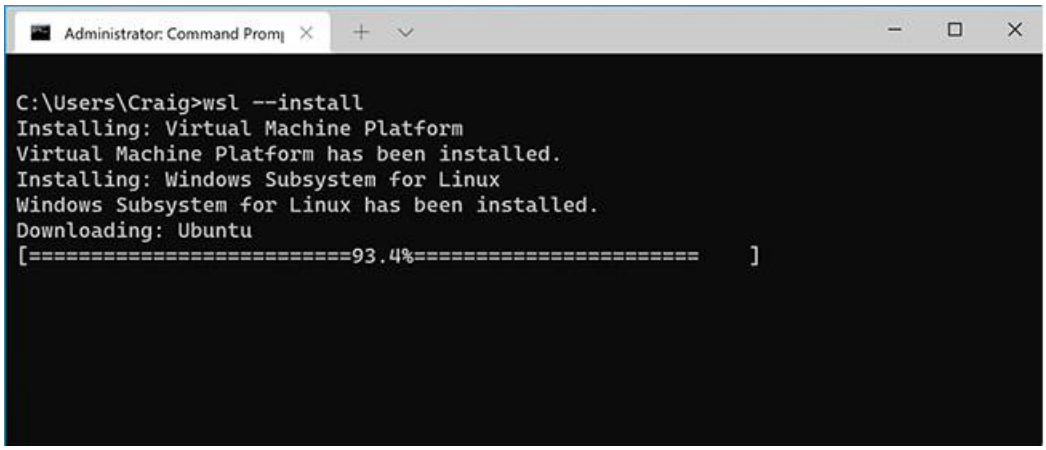
Kích hoạt **Virtual Machine Platform**

Bo mạch chủ và bộ xử lý phải hỗ trợ ảo hóa, đồng thời tùy chọn phải được bật trên **BIOS/UEFI.**

Tiếp theo mở cửa sổ CMD với quyền Admin và gõ lệnh:

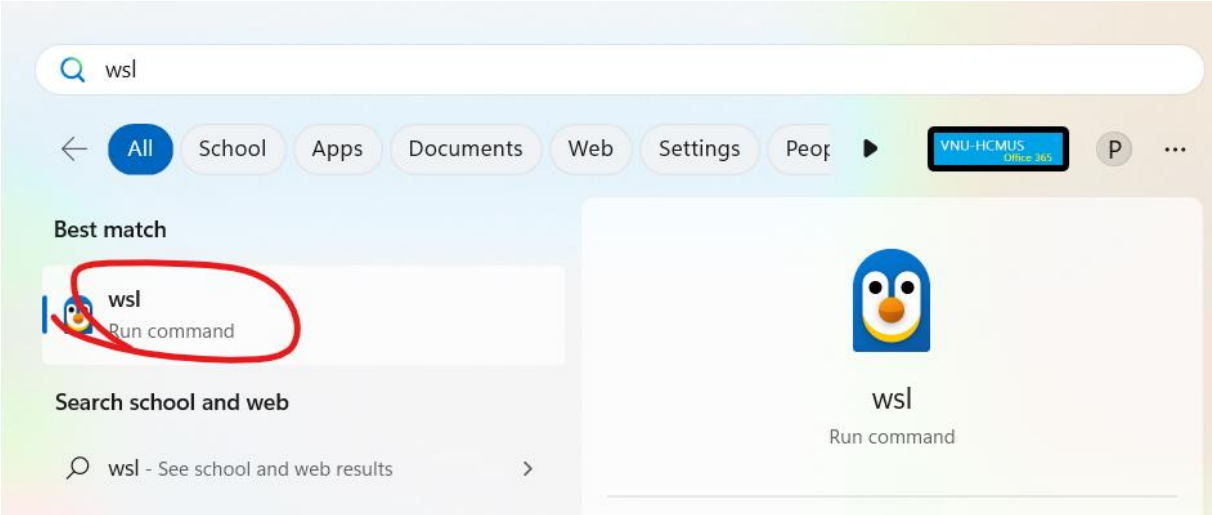
```
wsl.exe --install
```

Hệ thống sẽ tự động cài Ubuntu



Sau đó dùng lệnh **wsl –update** để cập nhật bản wsl mới nhất

Sau khi cài đặt đủ cho môi trường Window, tiến hành cài đặt Air Flow các bước sau:
Mở wsl và làm theo các bước sau:



Bước 1: Thiết lập môi trường ảo
Để làm việc với Airflow trên Windows, bạn cần thiết lập một môi trường ảo. Để làm điều này, bạn cần cài đặt gói virtualenv.

Lưu ý: Đảm bảo bạn đang ở gốc của cửa sổ dòng lệnh bằng cách gõ:

```
cd ~
pip install virtualenv
```

Tạo môi trường ảo như sau:

```
virtualenv airflow_env
```

Và sau đó kích hoạt môi trường:

```
source airflow_env/bin/activate
```

Tạo thư mục airflow và di chuyển đến thư mục vừa tạo.

Bước 2: Cài đặt Apache Airflow

Với môi trường ảo vẫn hoạt động và thư mục hiện tại đang trỏ đến thư mục Airflow đã tạo, cài đặt Apache Airflow:

```
pip install apache-airflow
```

Khởi tạo cơ sở dữ liệu:

```
airflow db init
```

Tạo một thư mục có tên là **dags** bên trong thư mục airflow. Thư mục này sẽ được sử dụng để lưu trữ tất cả các tập lệnh Airflow.

Bước 3: Tạo người dùng Airflow

Khi Airflow được cài đặt mới, bạn cần tạo một người dùng. Người dùng này sẽ được sử dụng để đăng nhập vào giao diện người dùng Airflow và thực hiện một số chức năng quản trị.

```
airflow users create --username admin --firstname admin --
lastname admin --role Admin --email youremail@email.com
```

Lưu ý: Nhớ đặt mật khẩu, tài khoản này được dùng truy cập trang quản lý hệ thống của airflow.
Kiểm tra người dùng đã tạo:

```
airflow users list
```

Bước 5: Chạy Webserver

```
airflow scheduler
```

Tạo một cửa sổ wsl thứ 2 và start web server bằng cách gõ lệnh sau:

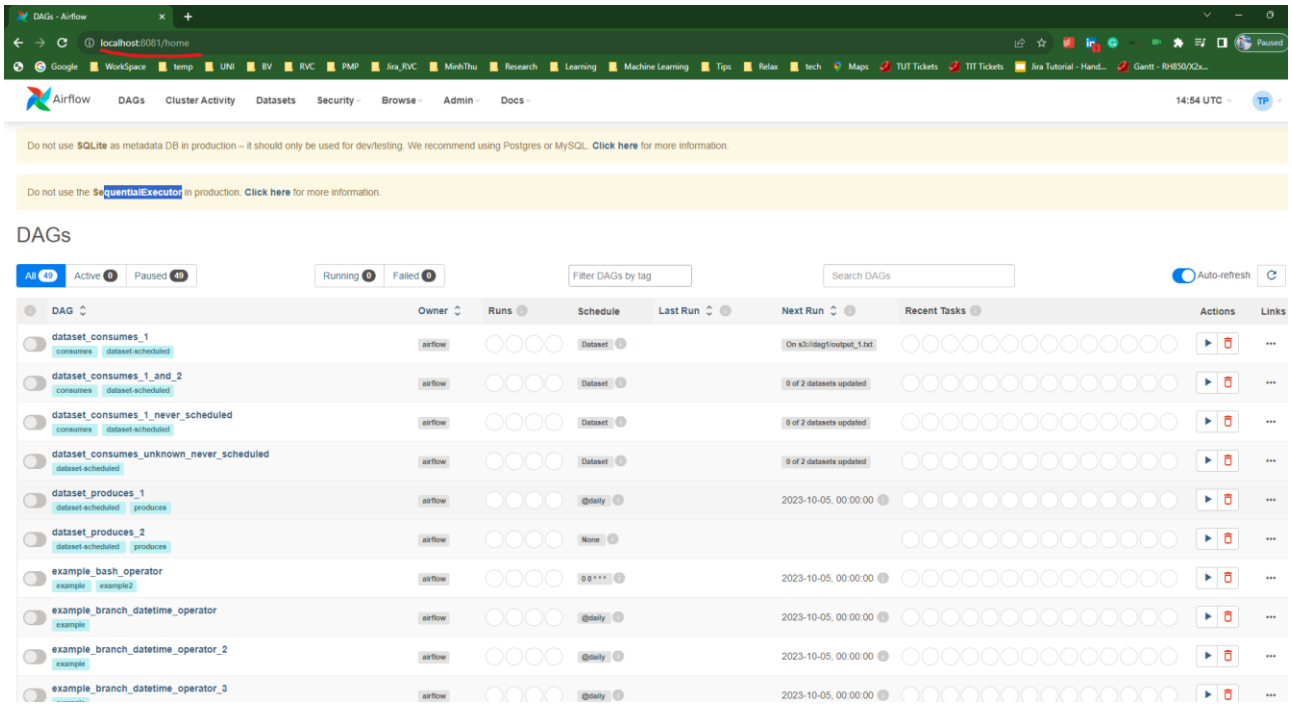
```
airflow webserver
```

Mặc định port là 8080 cho ứng dụng web quản lý.

Để thay đổi port ta dùng lệnh:

```
airflow webserver --port <port number>
```

Cuối cùng truy cập http://localhost:8081/home



IV. Tài liệu tham khảo

V. Bài tập