**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**KHOA TOÁN – TIN HỌC**

**CHUYÊN NGÀNH KHOA HỌC DỮ LIỆU**

**BÁO CÁO CUỐI KỲ:**

**KHAI THÁC DỮ LIỆU TỪ   
HỆ THỐNG WORLDBANK**

**MÔN HỌC:** KỸ THUẬT XỬ LÝ DỮ LIỆU

**GIẢNG VIÊN: TS. NGUYỄN THANH BÌNH**

**NHÓM 8:**

|  |  |  |
| --- | --- | --- |
| STT | TÊN | MSHV |
| 1 | Nguyễn Thị Kim Hoàng | 20C29022 |

Thành phố Hồ Chí Minh – 2021

1. **GIỚI THIỆU VỀ ĐỀ TÀI**

Với mục tiêu thực hiện được một quy trình đầy đủ để khai thác dữ liệu, em đưa ra một ví dụ thực nghiệm là: Khai thác dữ liệu từ trang web của Worldbank (<https://data.worldbank.org/indicator>).

Quy trình xử lý dữ liệu gồm nhiều bước cùng với sự lựa chọn các công cụ hỗ trợ, sau đây là sơ đồ thực hiện:

A picture containing text

Description automatically generated

Extract

Icon

Description automatically generated

Airflow

Load data lên Influxdb

Transform:

Unzip => Clean => Filtering

Truy vấn SQL và hiển thị trên Influxdb

Báo cáo gồm các nội dung:

1. Giới thiệu web <https://data.worldbank.org/indicator> và các dữ liệu.
2. Scraping dữ liệu
3. Quy trình biến đổi và load dữ liệu
4. Sử dụng Airflow
5. Truy vấn SQL và hiển thị dữ liệu bằng Influxdb
6. **GIỚI THIỆU WEB WORLDBANK VÀ CÁC DỮ LIỆU**

Graphical user interface, text, application, email

Description automatically generated

Các dữ liệu mà Ngân hàng Thế Giới cung cấp được thu thập trong vòng 50 năm qua, gồm nhiều chủ đề như tài chính, kinh doanh, y tế, kinh tế, phát triển con người. Hiện người dùng đã có thể truy cập trên 7.000 chỉ số phát triển.

Trang Chỉ số (Indicators) liệt kê 331 chỉ số trong bộ Chỉ số Phát triển Thế giới (WDI) theo vần ABC. Hàng năm, Ngân hàng Thế giới tổng hợp dữ liệu phát triển từ các dữ liệu gốc của mình và các nguồn khác được công nhận trên toàn thế giới để xây dựng bộ Chỉ số Phát triển Thế giới. Đây là công cụ để đánh giá tiến độ phát triển các nền kinh tế.

Các con số kể những câu chuyện về con người ở  các quốc gia mới nổi và đang phát triển, qua đó đóng góp một phần vào công cuộc xóa đói giảm nghèo.

Ở đồ án cuối kỳ này, tôi sẽ lấy các dữ liệu của trang Chỉ số này và dữ liệu được sử dụng để phân tích cũng như hiển thị trên một chương trình ở dưới máy tính cá nhân.

1. **QUY TRÌNH ETL**
   1. **Extract**

Việc lấy dữ liệu từ web được thực hiện bằng một đoạn code Python tên scraping.py, sử dụng thư viện Urllib và Xpath. Dữ liệu được lưu ở folder trên máy.

Text

Description automatically generated

* 1. **Transform - Load**

Quá trình transform - load gồm có 3 bước:

**Upzip => Cleaning => Processing**

Dữ liệu down về là các file .zip được chứa theo từng thư mục theo từng chỉ số:

Text

Description automatically generated

Do đó, dữ liệu cần đươc unzip. Việc unzip cũng được thực hiện bằng một đoạn code Python tên unzip.py.

Text

Description automatically generated

Với mỗi chỉ số tương ứng với từng file .zip, sau khi giải nén sẽ có 3 file .csv gồm một file chứa dữ liệu và 2 file metadata. File thực sự sử dụng chỉ gồm file có tên bắt đầu bằng API\_\*.csv. Ta lọc các file này ra bằng đoạn code Python với tên cleaning.py.

Text

Description automatically generated

Sau đây là một file dữ liệu điển hình mà ta có được:

Text

Description automatically generated

Tiếp đến ta cần xử lý file này để lấy ra các dữ liệu gồm:

* Last Update Date
* Country Name
* Country Code
* Indicator Name
* Indicator Code
* Years
* Values

Với bảng dữ liệu này, ta xử lý và load lên database của Influxdb bằng đoạn code Python với tên upload.py. Dữ liệu được định dạng theo yêu cầu riêng của Influxdb.

Text

Description automatically generated

1. **SỬ DỤNG AIRFLOW**

Các source code Python có thể run độc lập. Tuy nhiên, trong quá trình khai thác dữ liệu từ web, dữ liệu cần được cập nhật thường xuyên, các task được chạy tuần tự hoặc song song, việc thiết lập thời gian chạy của mỗi task thế nào thì cần được lên lịch trình và setup chạy cụ thể. Một ứng dụng hữu ích cho mục tiêu này đó là airflow.

Tất cả các task xây dựng đều có thể chạy song song. Nhưng vì nhiệm vụ Scraping, đây là task cần nhiều thời gian để hoàn thành. Ta để Scraping ở một DAG riêng lẻ. Việc download được setup scheduler là 1 ngày.

Các task còn lại sẽ được đưa vào DAG thứ hai và thử cách chạy tuần tự, scheduler được setup lặp lại theo từng phút. Sau mỗi task, dữ liệu ở thư mục input sẽ bị xóa đi để tránh chồng lắp dữ liệu.

Do đó, khi hoàn thành các task, các folder sẽ được tạo mới, nhưng đây là các folder trống.

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Các task được xây dựng bằng đoạn code Python được đặt trong folder dags của thư mục Airflow:

Text

Description automatically generated

1. **TRUY VẤN SQL VÀ HIỂN THỊ DỮ LIỆU BẰNG INFLUXDB**

InfluxDB là một cơ sở dữ liệu chuỗi thời gian phổ biến. Báo cáo sử dụng phiên bản Influxdb 1.8.9. Việc truy vấn trên Influxdb sử dụng một ngôn ngữ giống như SQL. Hiện nay, Influxdb đã phát triển đến phiên bản 2.0 và việc truy vấn đã chuyển sang ngôn ngữ NoSQL. Tuy nhiên, định dạng dữ liệu đưa vô chương trình là không thay đổi.

Graphical user interface, text, application, chat or text message

Description automatically generated

So với các hệ cơ sở dữ liệu khác, InfluxDB không cung cấp nhiều dạng biểu đồ, chủ yếu phục vụ cho hiển thị chuỗi dữ liệu thời gian, gồm có:

Graphical user interface, application

Description automatically generated

Đây là giao diện chương trình để lựa chọn dữ liệu hiển thị

A screenshot of a computer

Description automatically generated with medium confidence

Sau đây là một Dashboard ví dụ từ dữ liệu trang web WorldBank:

A screenshot of a computer

Description automatically generated with medium confidence

1. **KẾT LUẬN**

Tới đây, quá trình xử lý dữ liệu có thể xem như hoàn thành. Ta có thể phân tích dữ liệu một cách chi tiết hơn nữa. Việc lựa chọn Influxdb ở trong bài báo này không quá phù hợp. Influxdb rất phù hợp nhận dữ liệu streaming từ các thiết bị iot, hiển thị dữ liệu time series, đánh giá tốc độ hệ thống. Với dữ liệu ở báo cáo này, ta có thể tìm kiếm một hệ cơ sở dữ liệu tốt hơn để hiển thị.