Practice Project

Overview

Bối cảnh

Nhóm của bạn đã giao cho bạn nhiệm vụ tạo một quy trình Trích xuất, Chuyển đổi, Tải (ETL) tự động để trích xuất dự báo thời tiết hàng ngày và dữ liệu thời tiết quan sát được, sau đó tải chúng vào một báo cáo trực tiếp để nhóm phân tích sử dụng cho việc phân tích sâu hơn. Là một phần của dự án mô hình hóa dự báo lớn hơn, nhóm muốn sử dụng báo cáo này để theo dõi và đo lường độ chính xác lịch sử của dự báo nhiệt độ theo nguồn và trạm.

Để chứng minh khái niệm (proof-of-concept, POC), ban đầu bạn chỉ cần thực hiện việc này cho một trạm và một nguồn duy nhất. Đối với mỗi ngày vào buổi trưa (giờ địa phương), bạn sẽ thu thập cả nhiệt độ thực tế và nhiệt độ dự báo vào buổi trưa ngày hôm sau tại Casablanca, Morocco.

Trong giai đoạn sau, nhóm dự kiến sẽ mở rộng báo cáo để bao gồm danh sách các địa điểm, các nguồn dự báo khác nhau, tần suất cập nhật khác nhau và các số liệu thời tiết khác như tốc độ và hướng gió, lượng mưa và tầm nhìn.

Nguồn dữ liệu

Trong dự án thực hành này, bạn sẽ sử dụng gói dữ liệu thời tiết được cung cấp bởi dự án nguồn mở wttr.in, một dịch vụ web cung cấp thông tin dự báo thời tiết dưới dạng văn bản đơn giản.

Trước tiên, bạn sẽ sử dụng lệnh curl để thu thập dữ liệu thời tiết thông qua trang web wttr.in. Ví dụ: để lấy dữ liệu cho Casablanca, hãy nhập:

```
curl wttr.in/casablanca
```

Mục tiêu

- Tải xuống dữ liệu thời tiết thô
- Trích xuất dữ liệu quan tâm từ dữ liệu thô (
- Chuyển đổi dữ liệu theo yêu cầu

- Tải dữ liệu vào tệp nhật ký theo định dạng bảng
- Lên lịch để toàn bộ quy trình chạy tự động vào một thời điểm đã đặt hàng ngày

Các bước thực hiện

Bước 1: Khởi tạo

- Tạo file log report

```
touch project_log.log
```

- Khởi tạo header trong file report

```
echo -e "year\tmonth\tday\tobs_temp\tfc_temp">project_log.log
```

- Tạo file bash và cấp quyền

```
touch rx_poc.sh
#! /bin/bash
chmod u+x rx_poc.sh
```

Bước 2: Lấy dữ liệu thô từ **Casablanca**

Sử dụng lệnh curl với tùy chọn --output. Lưu kết quả vào tệp có tên weather_report.

```
city=Casablanca
curl -s wttr.in/$city?T --output weather_report
```

Bước 3: Trích xuất dữ liệu

3.1 Trích xuất nhiệt độ hiện tại

```
#To extract Current Temperature
obs_temp=$(curl -s wttr.in/$city?T | grep -m 1 '°.' | grep -Eo -e '-?[[:digit:]].
```

```
*')
echo "The current Temperature of $city: $obs_temp"
```

Giải thích:

- `curl -s wttr.in/$city?T`:
  - Gọi đến trang wttr.in, truy cập vào đường dẫn của thành phố `$city`
  - Thêm `?T` để lấy bản rút gọn (chỉ text, không màu).
  - `s` để curl chạy ở chế độ "silent", không in tiến trình.
- `grep -m 1 '°.'`:
  - Tìm dòng đầu tiên (`m 1`) có ký tự `°` (độ C hoặc F).
  - Mục đích là lấy nhiệt độ đầu tiên xuất hiện.
- `grep -Eo -e '-?[[:digit:]].*'`:
  - `E` dùng regex mở rộng.
  - `o` chỉ in phần khớp regex.
  - Biểu thức `?[[:digit:]].*` nghĩa là:
    - `?` : có thể có dấu (nhiệt độ âm).
    - `[[:digit:]]` : một chữ số.
    - `.*` : phần còn lại (tức là toàn bộ số + ký hiệu °C).
- Kết quả gán vào biến obs_temp

3.2 Trích xuất dữ liệu dự đoán cho chiều mai

```
# To extract the forecast tempearature for noon tomorrow
fc_temp=$(curl -s wttr.in/$city?T | head -23 | tail -1 | grep '°.' | cut -d 'C' -f2 | grep -Eo -e '-?[[:digit:]].*')
echo "The forecasted temperature for noon tomorrow for $city : $fc_temp C"
```

Giải thích:

- Gọi đến trang wttr.in, truy cập vào đường dẫn của thành phố `$city`
- `head -23`: giữ 23 dòng đầu tiên, `tail -1`: lấy dòng thứ 23 (Dòng thứ 23 trong output text thường tương ứng với nhiệt độ dự báo **giờ trưa ngày mai)**

- `grep '°'`: giữ lại phần có ký hiệu nhiệt độ (°)
- `cut -d 'C' -f2`: Cắt chuỗi theo ký tự **C** và giữ **phần sau** ký tự `C` (trong một số format của wttr.in thì nhiệt độ dự báo nằm sau ký tự C).
- `grep -Eo -e '-?[[:digit:]].*'`: Trích xuất chính xác phần số:
  - `?` → có thể có dấu âm.
  - `[[:digit:]]` → ít nhất một chữ số.
  - `.*` → phần còn lại (thường sẽ là `°`).
- Kết quả được gán vào biến fc_temp

3.3 Trích xuất các giá trị khác (Ngày, tháng, năm)

- Lấy dữ liệu

```
#Assign Country and City to variable TZ
TZ='Morocco/Casablanca'

# Use command substitution to store the current day, month, and year in co
rresponding shell variables:
day=$(TZ='Morocco/Casablanca' date -u +%d)
month=$(TZ='Morocco/Casablanca' date +%m)
year=$(TZ='Morocco/Casablanca' date +%Y)
```

- Thêm vào file log

```
record=$(echo -e "$year\t$month\t$day\t$obs_temp\t$fc_temp C")
echo $record>>project_log.log
```

Bước 4: Lập lịch để chạy script mỗi ngày

4.1 Xác định thời gian chạy script

```
date
Sat Aug 16 13:52:30 EDT 2025
date -u
Sat Aug 16 17:52:41 UTC 2025
```

Có thể thấy rằng múi h hiện tại là UTC + 4, và *Casablanca* nằm ở UTC+1, nên cần chạy script sớm hơn 3 tiếng so với buổi chiều ở chỗ hiện tại, là khoảng

4.2 Tạo cron

```
crontab -e
0 7 * * * /home/project/rx_poc.sh
```