# **Chương 1: Design**

1. **Data Source**

* API thời tiết được lấy từ [Weather Forecast](https://docs.tomorrow.io/reference/weather-forecast) . Đây là API trả về dự báo thời tiết cho 1 khoảng thời gian tương lai gần (phút/giờ/ngày) (tối đa 5 ngày hoặc 120 giờ).

1. **Dữ liệu trả về từ API:** Xem thêm trên tài liệu [Core](https://docs.tomorrow.io/reference/data-layers-core)  
   **Trong dự án: Sử dụng units = metric nên các đơn vị sẽ như bên dưới.**

* **time**: thời gian dự báo (theo giờ)
* **temperature**: Nhiệt độ đo được (Celsius [-90,60])
* **temperatureApparent**: Nhiệt độ cảm nhận bởi con người (Celsius [-90,60]), còn được hiểu là cột feels\_like\_temperature
* **dewPoint**: Nhiệt độ để hơi nước bắt đầu ngưng tụ thành giọt nước (Celsius [0,100])
* **humidity**: Độ ẩm không khí (%)
* **windSpeed**: tốc độ gió (m/s)
* **windDirection**: hướng gió (degrees)
* **windGust**: Gió giật - cơn gió mạnh nhất xảy ra thường dưới 20s (m/s)
* **pressureSurfaceLevel**: áp suất khí quyển tại mặt đất (hPa)
* pressureSeaLevel: áp suất khí quyển tại mặt biển (hPa)
* **precipitationIntensity**: Cường độ mưa/ tuyết, … tức thời tại mặt đất, thể hiện bằng lượng nước mưa sẽ rơi trong 1 giờ nếu tốc độ mưa giữ nguyên. (mm/hr)
* **rainIntensity:** Tốc độ mưa rơi tại thời điểm hiện tại, chỉ áp dụng cho mưa, không bao gồm tuyết hay các loại giáng thủy khác. (mm/hr)
* **freezingRainIntensity**
* **snowIntensity**
* **sleetIntensity:** Cường độ mưa tuyết đá, tương tự các cột trên
* **precipitationProbability:** Xác suất có mưa (hoặc tuyết...) tại một địa điểm, trong khoảng thời gian dự báo, nếu lượng giáng thủy đạt từ 0.254 mm trở lên. (%)
* **precipitationType:** Loại thời tiết có thể xảy ra

| Rain | Mưa |
| --- | --- |
| Snow |  |
| Freezing Rain | Dạng lỏng+đóng băng khi chạm đất |
| Ice Pellets/ Sleet | Mưa tuyết đá |

* **rainAccumulation**: Lượng mưa tích lũy – tổng lượng nước mưa dạng lỏng đã hoặc sẽ rơi trong 1 giờ tại thời điểm được yêu cầu. Đơn vị: mm hoặc inch
* **snowAccumulation**
* **snowAccumulationLwe**: Tuyết tích lũy quy đổi thành nước (trong 1 giờ), không cộng dồn theo ngày. (mm, in)
* **snowDepth:** Độ sâu của tuyết (cm, in)
* **sleetAccumulation:** (mm, in)
* **sleetAccumulationLwe:** Lượng nước tương đương của mưa tuyết đá tích lũy – lượng nước (mm/inch) thu được nếu toàn bộ sleet rơi trong 1 giờ tan chảy.
* **iceAccumulation:** chiều dày của lớp băng tích tụ được từ hiện tượng này trong 1 giờ. (mm, in)
* **iceAccumulationLwe:** (mm, in)
* **sunriseTime:** Thời điểm mặt trời mọc (vd: 2025-07-18T05:34:00Z)
* **sunsetTime:** Thời điểm mặt trời lặn (vd: 2025-07-18T05:34:00Z)
* **Visibility:** Tầm nhìn xa – khoảng cách tối đa mà một vật thể hoặc ánh sáng có thể được nhìn thấy rõ ràng trong điều kiện hiện tại. (km)
* **cloudCover:** Phần bầu trời bị mây che phủ, được tính theo tỷ lệ (%) tại một thời điểm và vị trí nhất định.
* **cloudBase:** Đáy mây – độ cao thấp nhất của phần mây có thể nhìn thấy được, tính từ mặt đất. Đơn vị: km
* **cloudCeling:** Trần mây – độ cao thấp nhất của lớp mây liên tục che phủ phần lớn bầu trời (≥ 5/8), tính từ mặt đất. Đơn vị: km
* **moonPhase:** thể hiện hình dạng phần Mặt Trăng được Mặt Trời chiếu sáng trực tiếp khi nhìn từ Trái Đất — hay còn gọi là pha Mặt Trăng.
  + "0": "New",
  + "1": "Waxing Crescent",
  + "2": "First Quarter",
  + "3": "Waxing Gibbous",
  + "4": "Full",
  + "5": "Waning Gibbous",
  + "6": "Third Quarter",
  + "7": "Waning Crescent"
* **uvIndex:** Đo mức độ cao thấp của tia UV 0-2: Low 3-5: Moderate 6-7: High 8-10: Very High 11+: Extreme
* **uvHealthConcern:** Mức cảnh báo sức khỏe tương đương với chỉ số tia UV (uvIndex) 0-2: Low 3-5: Moderate 6-7: High 8-10: Very High 11+: Extreme
* **gdd10To30:** chỉ số Growing Degree Days (GDD) — tức tích nhiệt độ phát triển cây trồng — với ngưỡng tối thiểu 10°C và tối đa 30°C.
* **gdd10To31**: chỉ số Growing Degree Days (GDD) — tức tổng nhiệt độ có ích cho sinh trưởng — với: tối thiểu là 10 độ C và tối đa là 31 độ C
* **gdd08To30:** chỉ số Growing Degree Days (GDD) — tích nhiệt sinh trưởng với tối thiểu là 8 độ C và tối đa là 30 độ C
* **gdd03To25:** chỉ số Growing Degree Days (GDD) – tích nhiệt độ có ích cho sự phát triển cây trồng, được tính với tối thiểu là 3 độ C và tối đa là 25 độ C

→ Các cột bắt đầu bằng GDD (Growing Degree Days) trong dữ liệu thời tiết được dùng để: **Đánh giá tiến độ phát triển của cây trồng theo từng giai đoạn sinh trưởng, dựa trên tổng nhiệt tích lũy mà cây nhận được.**

* **Evapotranspiration** (bốc – thoát hơi nước): Tổng lượng nước bị mất do bốc hơi từ mặt đất và thoát hơi từ cây trồng vào khí quyển, tính theo phương pháp Penman–Monteith FAO56.
* **weatherCodeFullDay:** Mô tả điều kiện thời tiết chính trong suốt cả ngày (từ bình minh hôm nay đến bình minh hôm sau).
* **weatherCodeDay:** Mô tả thời tiết chính trong ban ngày (từ lúc mặt trời mọc đến lúc mặt trời lặn).
* **weatherCodeNight:** Mô tả thời tiết chính trong ban đêm (từ lúc mặt trời lặn đến lúc mặt trời mọc ngày hôm sau).
* **weatherCode:** Mô tả thời tiết chính tại thời điểm hiện tại hoặc tại một thời điểm cụ thể trong dự báo.
* **thunderstormProbability**: Xác suất xảy ra giông sét – được tính theo khả năng mật độ sét đánh vượt quá 0.1 lần trong mỗi 100 km² mỗi giờ. (đơn vị %)
* **ezHeatStressIndex**: Chỉ số mức độ căng thẳng nhiệt (heat stress) dựa theo công thức Ezra-Zohar. Giá trị càng cao → cơ thể càng chịu nhiều áp lực do nhiệt độ và độ ẩm. 0–22: Không gây căng thẳng nhiệt / 22–24: Căng thẳng nhẹ / 24–26: Căng thẳng vừa / 26–28: Căng thẳng cao vừa / 28–30: Căng thẳng nghiêm trọng / 30: Căng thẳng nhiệt cực độ

1. **Mô tả dự án**

Đầu tiên sử dụng Airflow để lên lịch lấy dữ liệu tự động từ API này, sau đó trích xuất dữ liệu vào các files csv (stagging). Tiếp tục xử lý dữ liệu này với PySpark và load dữ liệu đã xử lý vào PostgreSQL. Cuối cùng là trực quan hóa và phân tích insight với Power BI.

1. **Các công nghệ sử dụng và lý do chọn**

**Apache Airflow:** Lên lịch tự động để lấy dữ liệu từ API thời tiết

**Apache Spark / PySpark:** Sử dụng để xử lý tối ưu cho dữ liệu lớn với cách phân tán công việc cluster computing,... thay vì Pandas chạy trên bộ nhớ RAM

**PostgreSQL:** Sử dụng xem như là data warehouse - lưu trữ dữ liệu đã được xử lý

**Power BI:** Phân tích trực quan hóa dữ liệu với các bài toán đặt ra

**Docker:** Đóng gói dự án thành container để các máy khác dễ dàng chạy, không cần tạo thư viện hay setup môi trường

**Github:** Lưu trữ mã nguồn của dự án

1. **Các bài toán đặt ra**

* Phân tích xu hướng nhiệt độ tại … trong ngắn hạn
* Trực quan hóa dữ liệu thời tiết
* Tự động thu thập và lưu trữ dữ liệu thời tiết

1. **Chiến lược gọi API**

**Loại dữ liệu:** Dự báo thời tiết (forecast)

**Granularity:** Hourly forecast

**Số địa điểm:** 10 địa điểm cố định

**Tần suất cập nhật:** Mỗi 6 giờ (có thể linh hoạt thành 3–12h tùy nhu cầu)

**Rate limit của API:**

* 1 giây: 3 lần
* 1 giờ: 25 lần
* 1 ngày: 500 lần

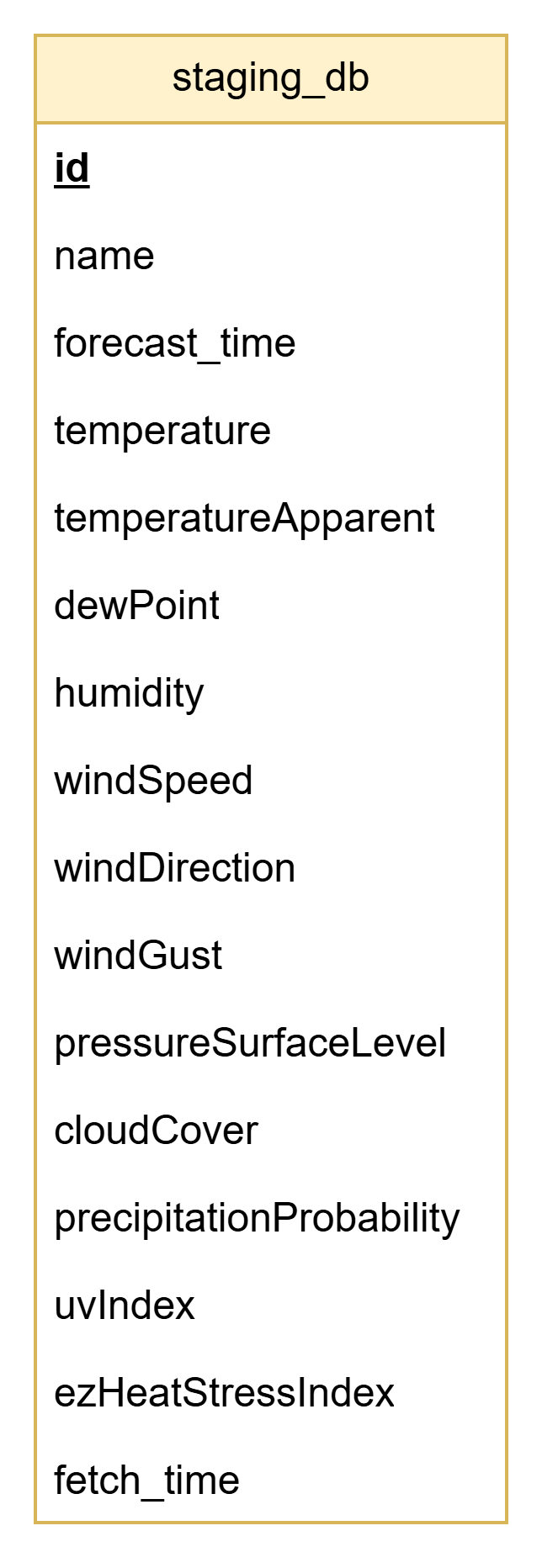
→ Mỗi 6h / lần

10 địa điểm = 10 requests

Mỗi requests cách nhau 0.5 giây

Dùng retry nếu status != 200 hoặc timeout

1. **Thiết kế database lưu trữ**

****

1. **Xác định các bài toán**
2. **Thiết kế hệ thống ETL**

****

1. **Taskflow**
2. **FIle system**

*└── weather-data-pipeline/*

*├── .env*

*├── .gitignore*

*├── docker-compose.yaml*

*├── README.md*

*├── requirements.txt*

*├── test.ipynb*

*├── weather\_data.csv*

*├── staging\_db/*

*│ └── staging\_weather\_db.sql*

*├── scripts/*

*│ └── extract.py*

*├── diagram/*

*│ ├── .$weather-diagram.drawio.bkp*

*│ └── weather-diagram.drawio*

*├── dags/*

*│ ├── weather\_dag.py*

*│ └── \_\_pycache\_\_/*

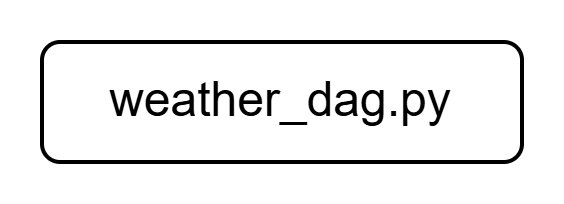
*│ ├── fetch\_weather\_data.cpython-312.pyc*

*│ ├── load\_to\_staging\_mysql.cpython-312.pyc*

*│ └── weather\_dag.cpython-312.pyc*

*└── logs/*

1. **Thiết kế DAGs**

****

# **Chương 2: Thực hiện**