



Đề tài: Dự báo thời tiết dựa trên Bigdata



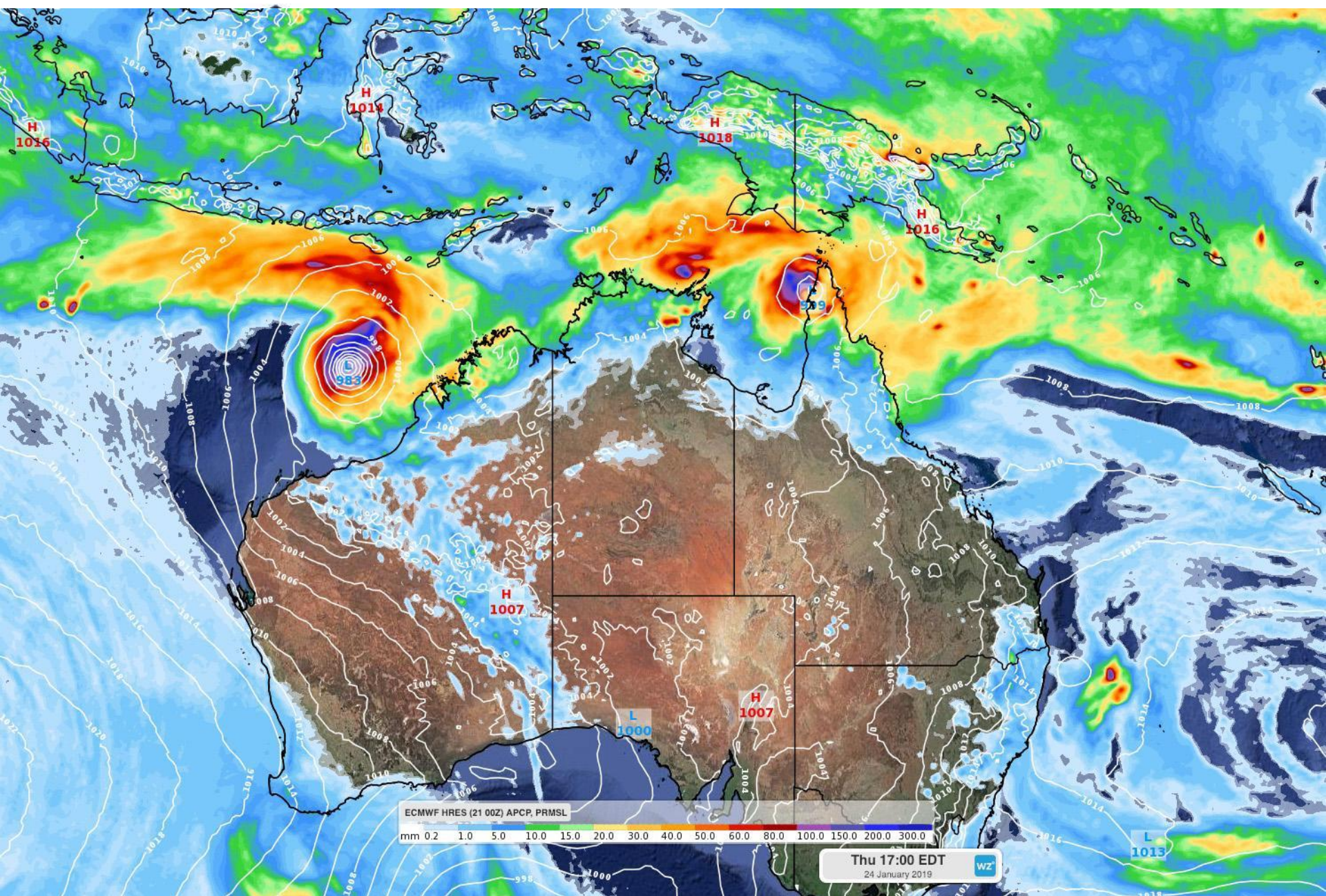
Thành viên (5)

1. Nguyễn Đỗ Tú - 20194200
2. Nguyễn Trung Kiên - 20194088
3. Nguyễn Lê Tài - 20194162
4. Nguyễn Văn Cường - 20194001
5. Nguyễn Văn Thịnh - 20194178



Phân công nhiệm vụ

Họ và tên	Công việc
Nguyễn Đỗ Tú	Thiết kế luồng và phân tích dữ liệu
Nguyễn Trung Kiên	Model dự báo thời tiết + Crawl data
Nguyễn Văn Cường	Model dự báo thời tiết
Nguyễn Văn Thịnh	Crawl data
Nguyễn Lê Tài	Model dự báo thời tiết





Mục lục

- I. Mô tả đề tài và dữ liệu
- II. Kiến trúc hệ thống



I. Mô tả đề tài

- Thu thập dữ liệu liên quan đến khí tượng thủy văn tại khu vực sân bay Nội Bài trên **wunderground**
- Xử lý và lưu trữ dữ liệu dựa trên airflow.
- Các kĩ thuật sử dụng để đưa ra dự báo về thời tiết: Deep Learning



I. Mô tả dữ liệu

- Format: Time, Temperature, Dew_point, Humidity, Wind, Wind_speed, Wind_gust, Presure, Presip, Condition

+ **Time**: Thời gian crawl data

+ **Temperature (°F)**: Nhiệt độ

+ **Dew_point (°F)**: nhiệt độ điểm sương, là điểm nhiệt độ mà khi đó không khí mát bắt đầu hòa làm một với hơi nước.

+ **Humidity (%)**: Độ ẩm

+ **Wind** (East, West, South, North): Hướng gió

+ **Wind_speed (mph)**: Tốc độ gió

+ **Wind_gust (mph)**: Tốc độ giật của gió

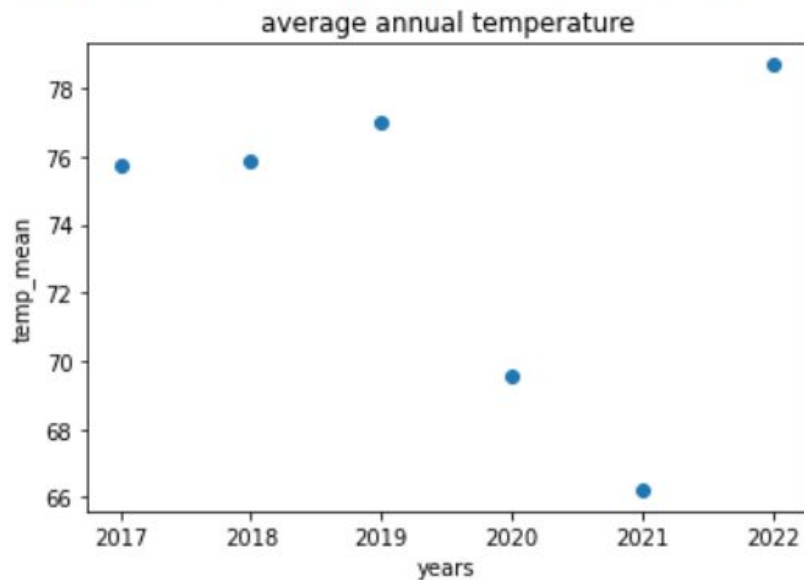
+ **Presure (in)**: Áp suất khí quyển

(At sea level, standard air pressure in inches of mercury is 29.92)

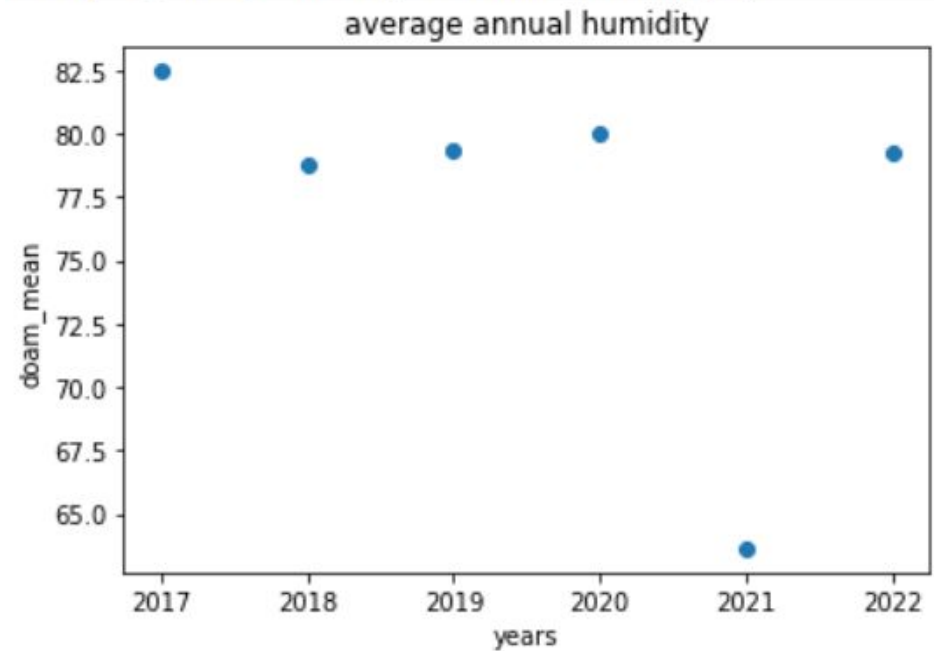
+ **Presip (in)**: giáng thủy

+ **Condition**: trạng thái thời tiết (VD: Sương mù-Fog, Nhiều mây-Mostly Cloud, ...)

I . Visualize dữ liệu



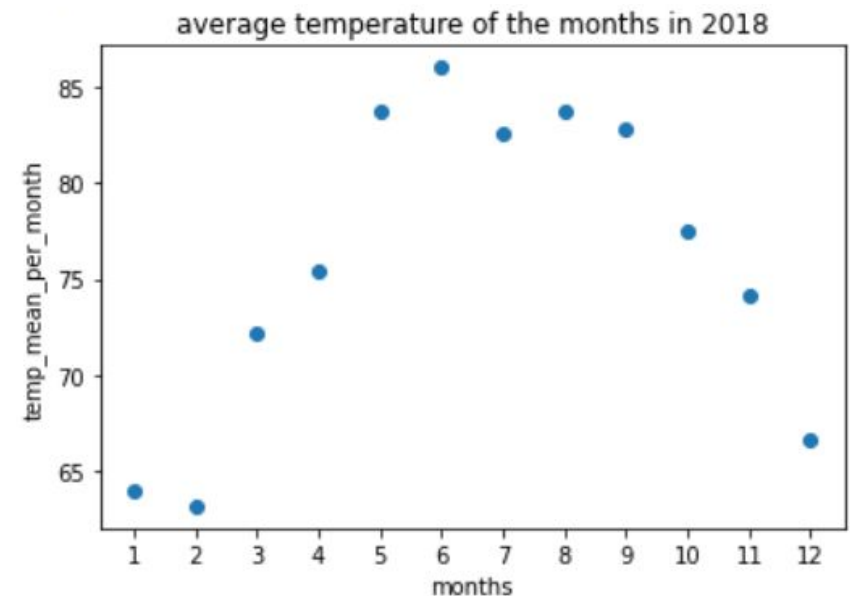
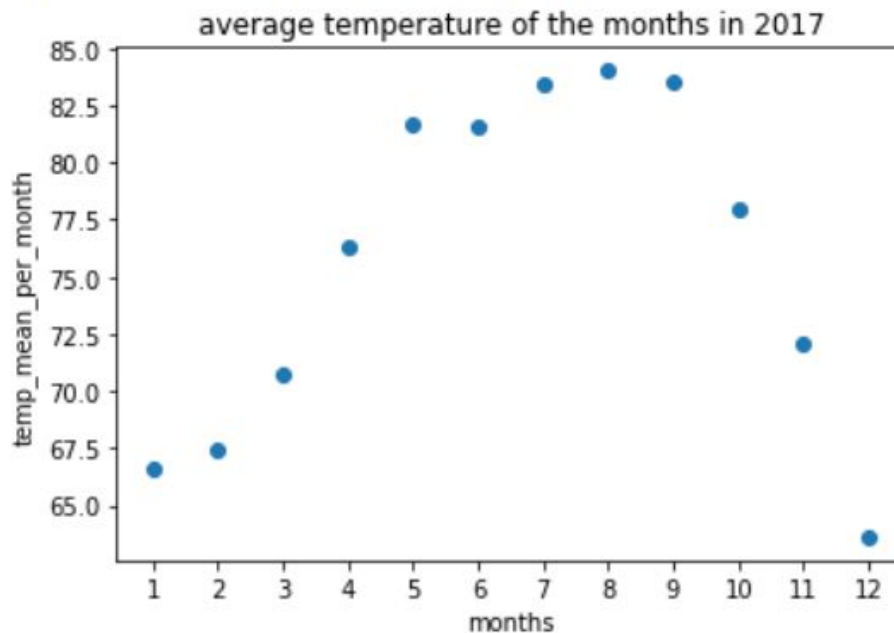
Biểu đồ so sánh giá trị nhiệt độ trung bình của các năm



Biểu đồ so sánh giá trị độ ẩm trung bình của các năm

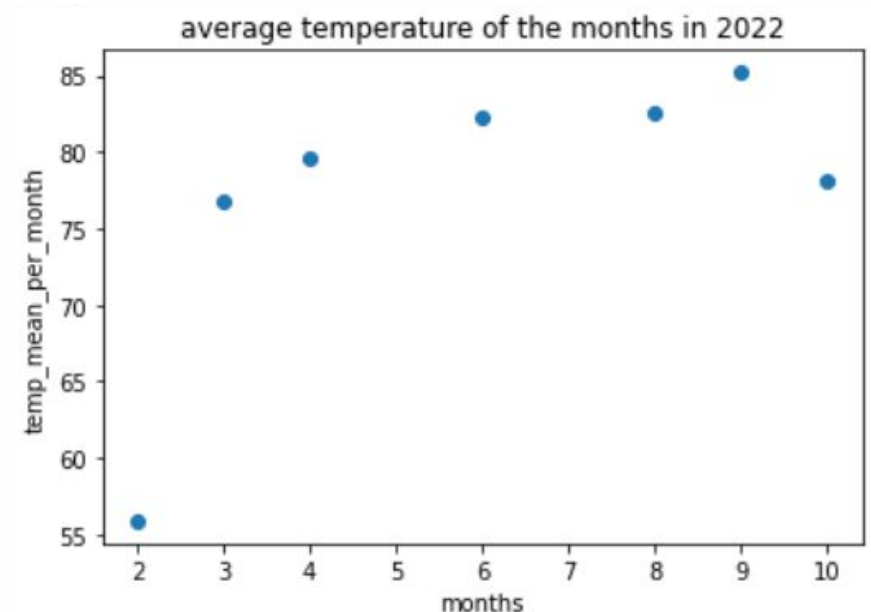
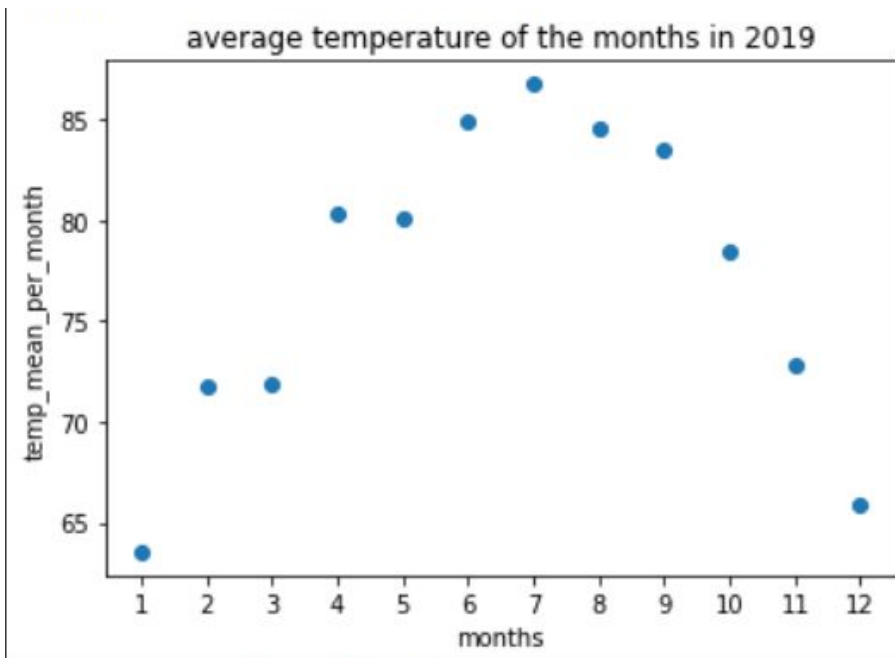
I . Visualize dữ liệu

Biểu đồ so sánh giá trị nhiệt độ trung bình của các tháng trong năm



I . Visualize dữ liệu

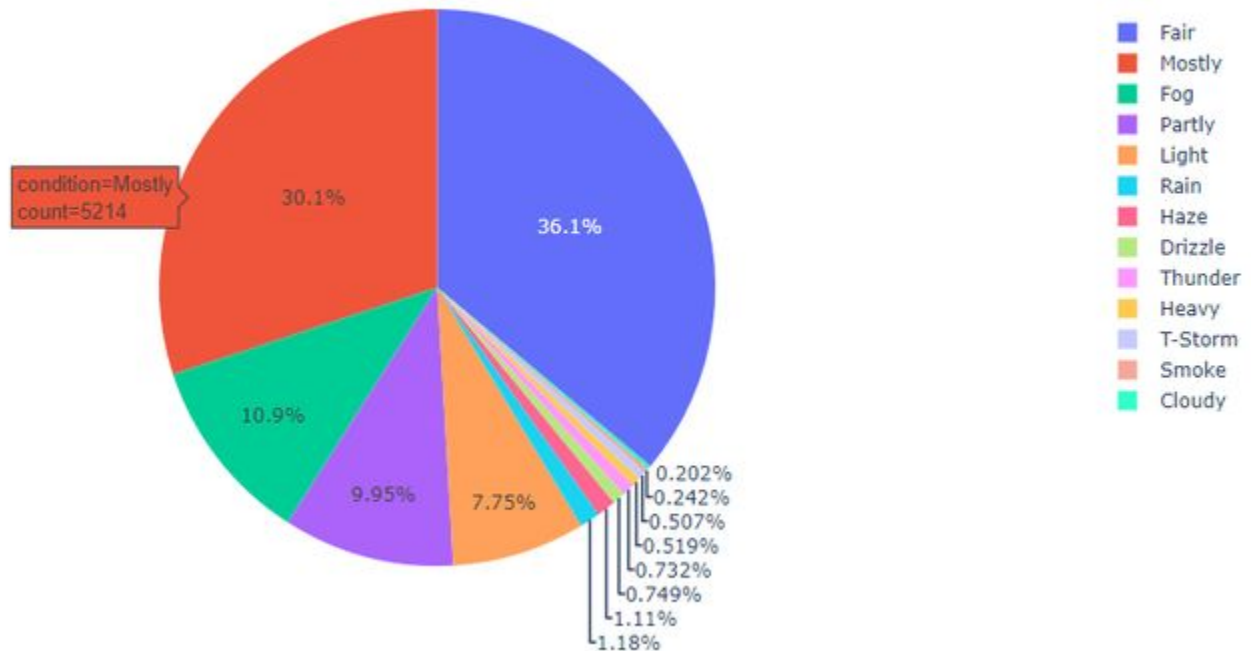
Biểu đồ so sánh giá trị nhiệt độ trung bình của các tháng trong năm



I . Visualize dữ liệu

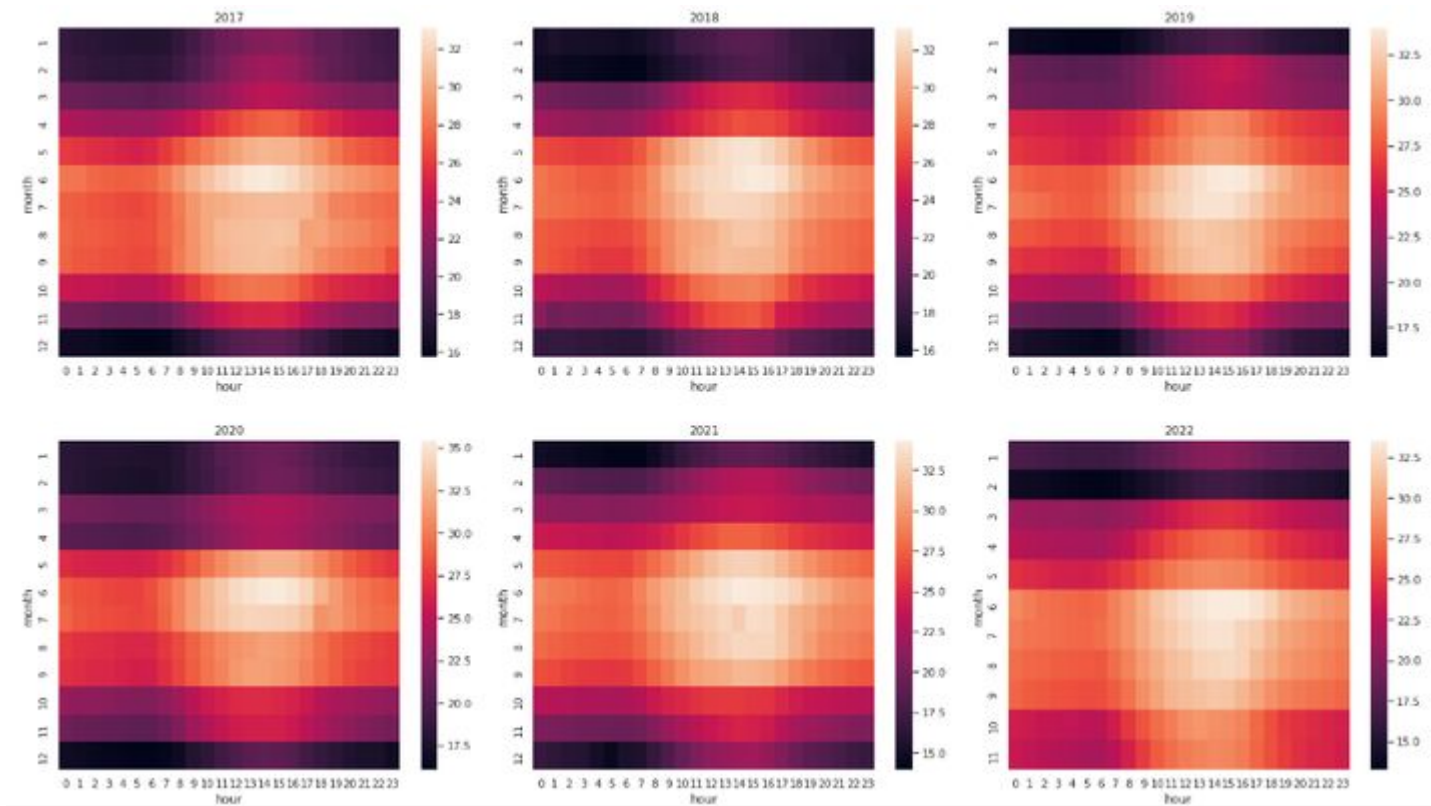
Phần trăm các điều kiện thời tiết xuất hiện trong năm

Percentage of conditions type in year



I . Visualize dữ liệu

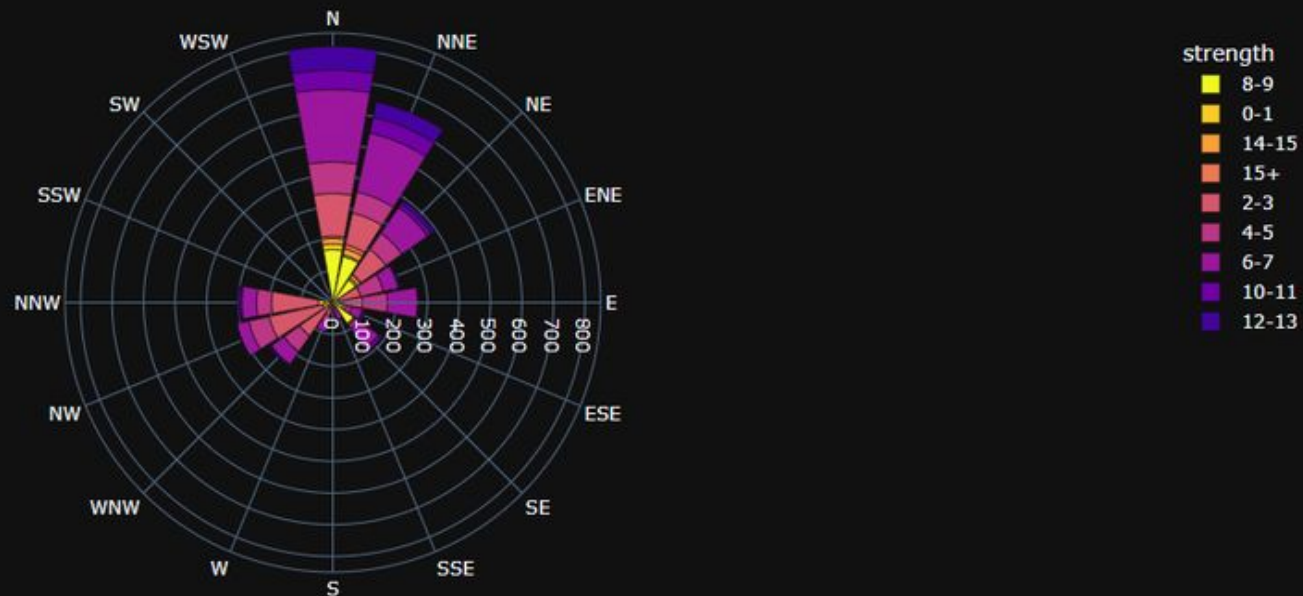
Nhiệt độ trung bình



I . Visualize dữ liệu

Hướng gió (mùa xuân và mùa đông)

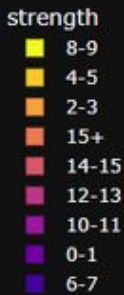
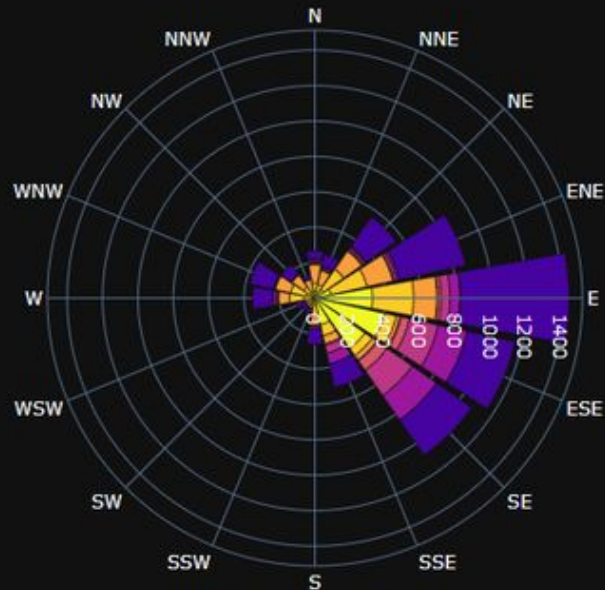
winter + spring



I . Visualize dữ liệu

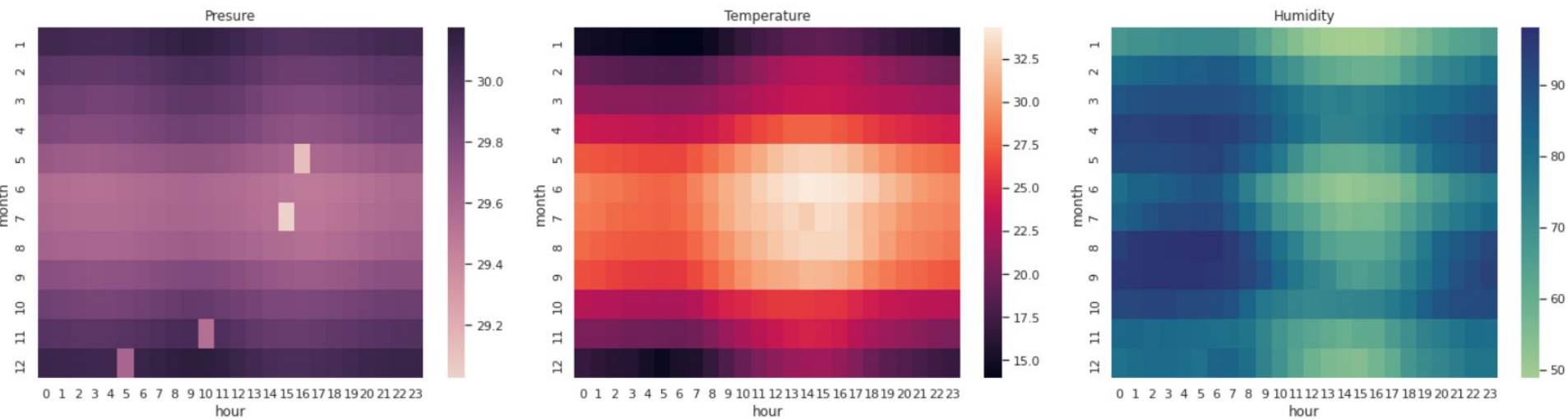
Hướng gió (mùa hạ và mùa thu)

summer + autumn



I . Visualize dữ liệu

Biểu đồ so sánh Áp suất - Nhiệt độ - Độ ẩm



I . Visualize dữ liệu

Biểu đồ tương quan giữa các trường dữ liệu



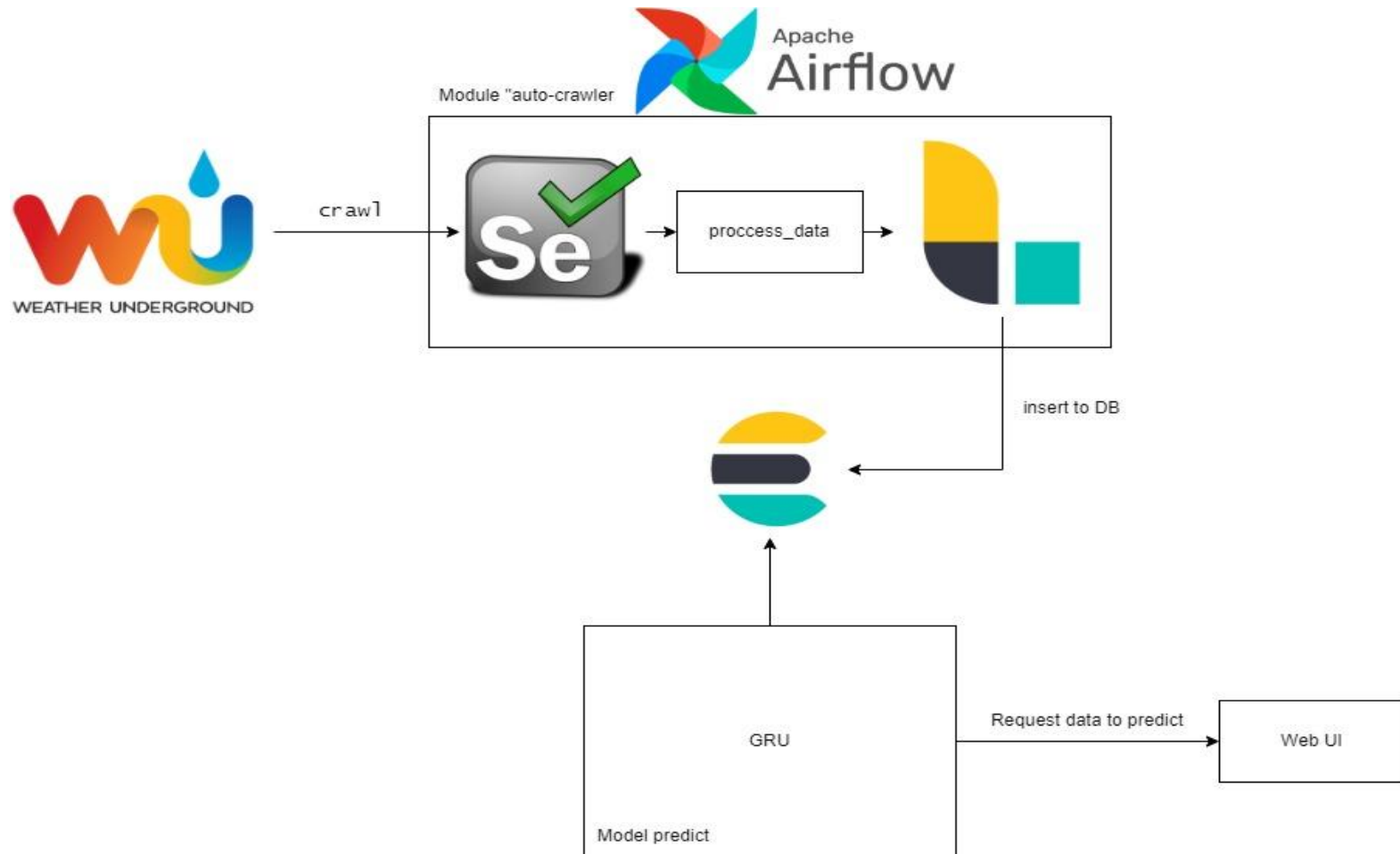


II. Kiến trúc hệ thống

- Thu thập dữ liệu liên quan đến khí tượng thủy văn tại khu vực sân bay Nội Bài trên **wunderground**
- Xử lý và lưu trữ dữ liệu dựa trên airflow.
- Các kĩ thuật sử dụng để đưa ra dự báo về thời tiết: Deep Learning

II. Kiến trúc hệ thống

Kiến trúc tổng thể hệ thống



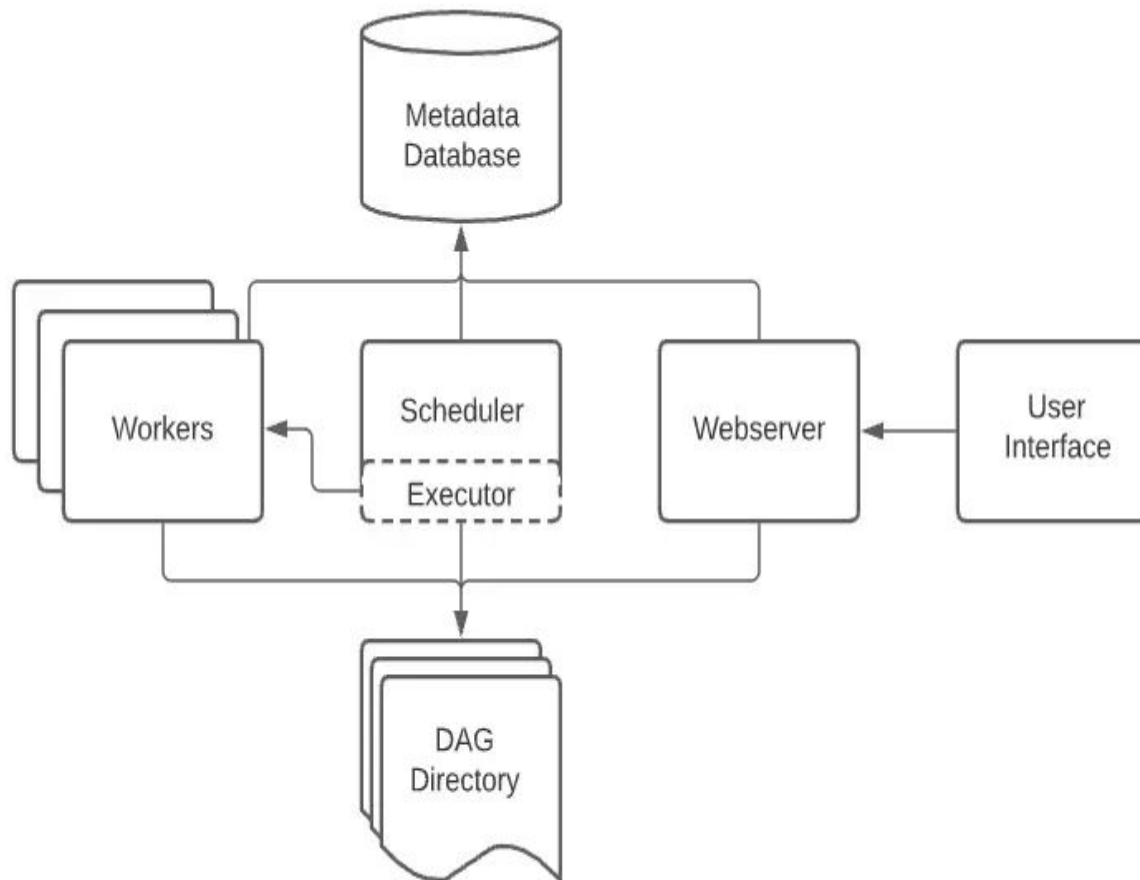


II. Kiến trúc hệ thống

Airflow là một công cụ lập lịch trình cho luồng công việc của bạn cũng như hỗ trợ quản lý, theo dõi từng phần trong quy trình giúp bạn sửa lỗi, bảo trì code thuận tiện và dễ dàng.



II. Kiến trúc hệ thống

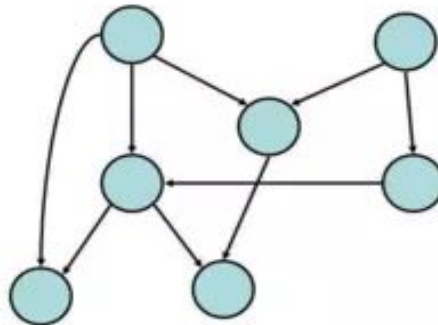


I. Kiến trúc hệ thống


- Workflow
 - Airflow có thể tự động hóa quy trình công việc bằng DAGs

Directed Acyclic Graph

- DAG – directed graph with no directed cycles

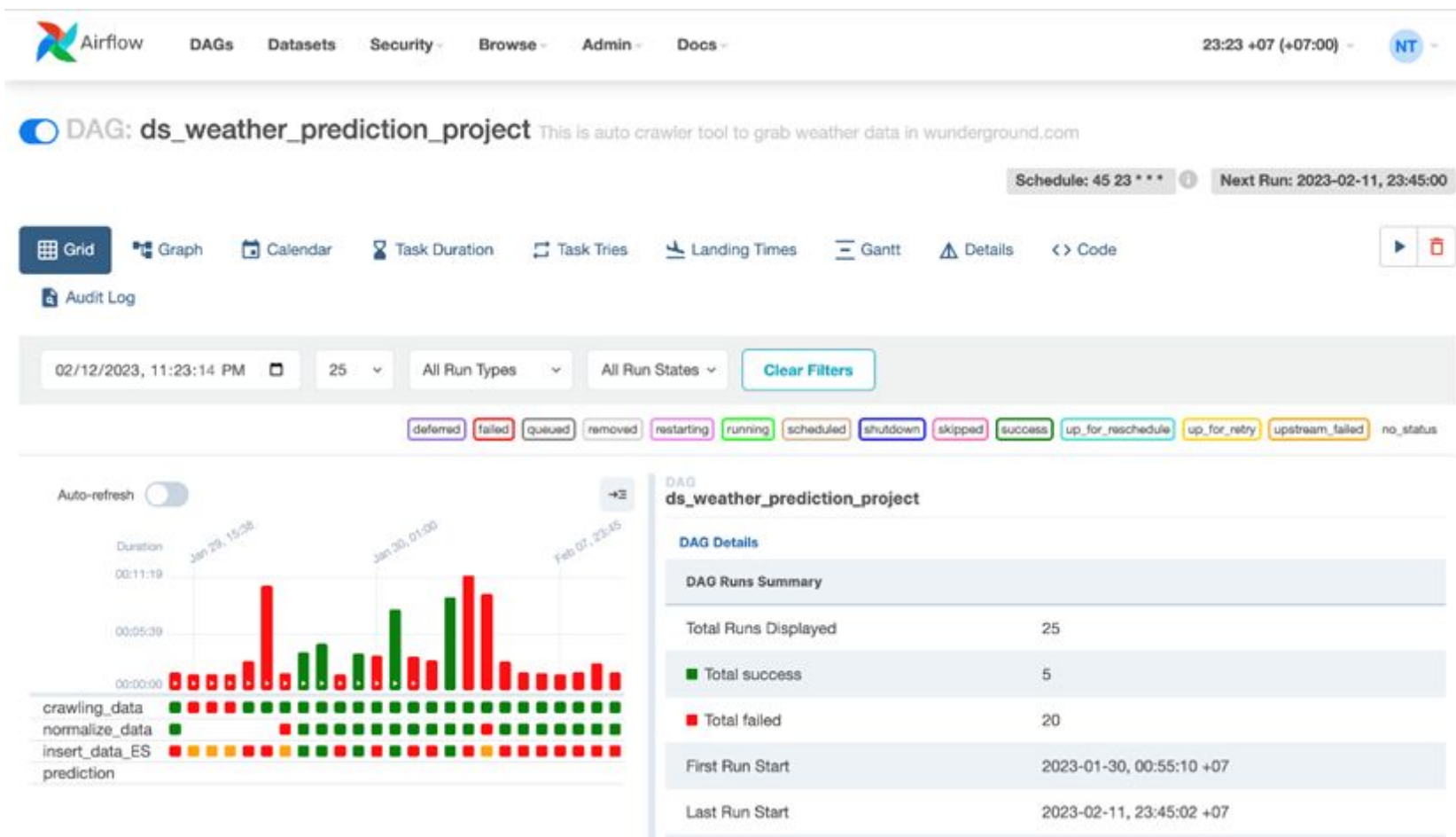


Airflow

 Airflow DAGs Datasets Security Browse Admin Docs 23:23 +07 (+07:00) NT							
DAGs							
All 44 Active 1 Paused 43		Filter DAGs by tag		Search DAGs		Auto-refresh	
DAG	Owner	Runs	Schedule	Last Run	Next Run	Recent Tasks	
<input type="checkbox"/> dataset_consumes_1 consumes dataset-scheduled	airflow	<div><div></div><div></div><div></div><div></div><div></div></div>	Dataset		On s3://dag1/output_1.txt	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	
<input type="checkbox"/> dataset_consumes_1_and_2 consumes dataset-scheduled	airflow	<div><div></div><div></div><div></div><div></div><div></div></div>	Dataset		0 of 2 datasets updated	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	
<input type="checkbox"/> dataset_consumes_1_never_scheduled consumes dataset-scheduled	airflow	<div><div></div><div></div><div></div><div></div><div></div></div>	Dataset		0 of 2 datasets updated	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	
<input type="checkbox"/> dataset_consumes_unknown_never_scheduled dataset-scheduled	airflow	<div><div></div><div></div><div></div><div></div><div></div></div>	Dataset		0 of 2 datasets updated	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	
<input type="checkbox"/> dataset_produces_1 dataset-scheduled produces	airflow	<div><div></div><div></div><div></div><div></div><div></div></div>	@daily		2023-02-11, 07:00:00	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	
<input type="checkbox"/> dataset_produces_2 dataset-scheduled produces	airflow	<div><div></div><div></div><div></div><div></div><div></div></div>	None			<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	
<input checked="" type="checkbox"/> ds_weather_prediction_project	lund	<div><div></div><div></div><div></div><div></div><div></div></div>	45 23 ***	2023-02-10, 23:45:00	2023-02-11, 23:45:00	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	
<input type="checkbox"/> example_bash_operator example example2	airflow	<div><div></div><div></div><div></div><div></div><div></div></div>	0 0 ***		2023-02-11, 07:00:00	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	
<input type="checkbox"/> example_branch_datetime_operator example	airflow	<div><div></div><div></div><div></div><div></div><div></div></div>	@daily		2023-02-11, 07:00:00	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	

Airflow UI: danh sách các DAGs đã được cài đặt

Airflow



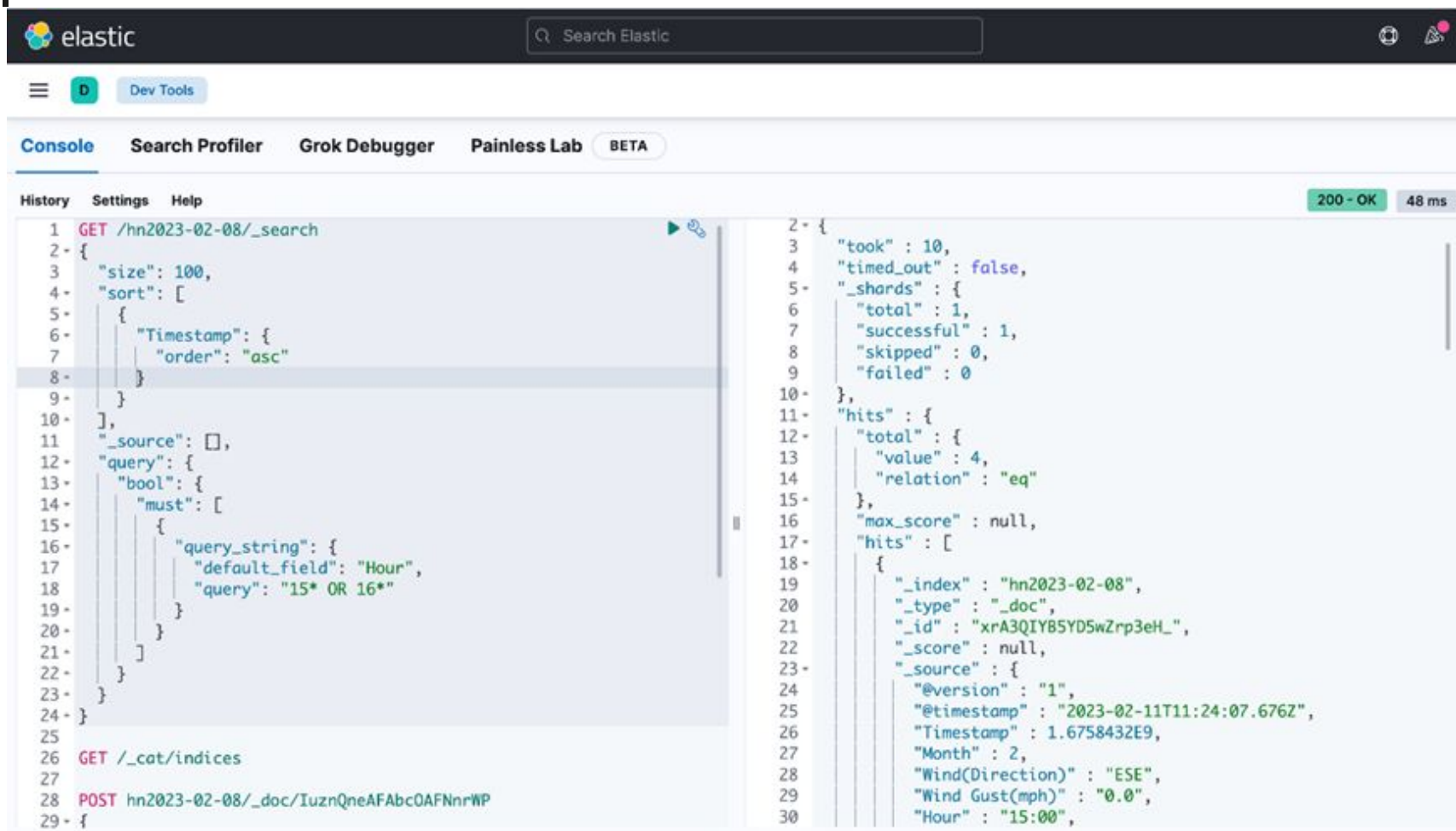
Airflow UI: chi tiết về một DAG



ELK stack

- **ELK: Elasticsearch + Logstash + Kibana**
 - Logstash tải nhập, chuyển đổi và gửi dữ liệu đến đúng điểm đích.
 - Elasticsearch lập chỉ mục, phân tích và tìm kiếm dữ liệu đã tải nhập.
 - Kibana hiển thị kết quả phân tích.
- **ELK Stack mang tới cho bạn khả năng tổng hợp nhật ký từ tất cả các hệ thống và ứng dụng của bạn, phân tích những nhật ký này, hiển thị dữ liệu để giám sát ứng dụng và cơ sở hạ tầng, khắc phục sự cố nhanh hơn, phân tích bảo mật, v.v.**

ELK stack



The screenshot displays the Kibana console interface. At the top, the Elastic logo and a search bar are visible. Below the navigation bar, the 'Console' tab is active. The console shows a series of commands and their results:

```
1 GET /hn2023-02-08/_search
2 {
3   "size": 100,
4   "sort": [
5     {
6       "Timestamp": {
7         "order": "asc"
8       }
9     }
10  ],
11  "_source": [],
12  "query": {
13    "bool": {
14      "must": [
15        {
16          "query_string": {
17            "default_field": "Hour",
18            "query": "15* OR 16*"
19          }
20        }
21      ]
22    }
23  }
24 }
```

The results of the search are displayed on the right side of the console:

```
2 {
3   "took": 10,
4   "timed_out": false,
5   "_shards": {
6     "total": 1,
7     "successful": 1,
8     "skipped": 0,
9     "failed": 0
10  },
11  "hits": {
12    "total": {
13      "value": 4,
14      "relation": "eq"
15    },
16    "max_score": null,
17    "hits": [
18      {
19        "_index": "hn2023-02-08",
20        "_type": "_doc",
21        "_id": "xrA3QIY85YD5wZrp3eH_",
22        "_score": null,
23        "_source": {
24          "@version": "1",
25          "@timestamp": "2023-02-11T11:24:07.676Z",
26          "Timestamp": 1.6758432E9,
27          "Month": 2,
28          "Wind(Direction)": "ESE",
29          "Wind Gust(mph)": "0.0",
30          "Hour": "15:00",

```

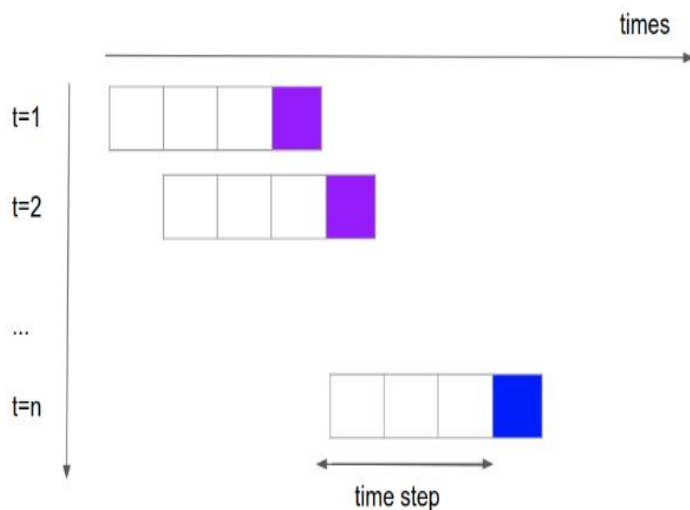
The console also shows the following commands and their results:

```
26 GET /_cat/indices
27
28 POST hn2023-02-08/_doc/IuznQneAFAbcOAFNnrWP
29 {
```

Kibana: giao diện trực quan hóa Elasticsearch

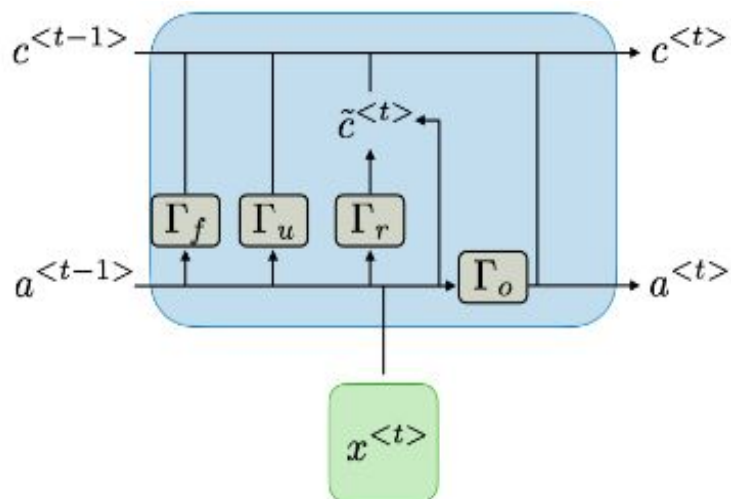
II . Kiến trúc hệ thống - mô hình

- Giả định của các mô hình chuỗi thời gian là các qui luật trong quá khứ sẽ được lặp lại ở tương lai nên thông thường mô hình được huấn luyện trên những dữ liệu trong quá khứ



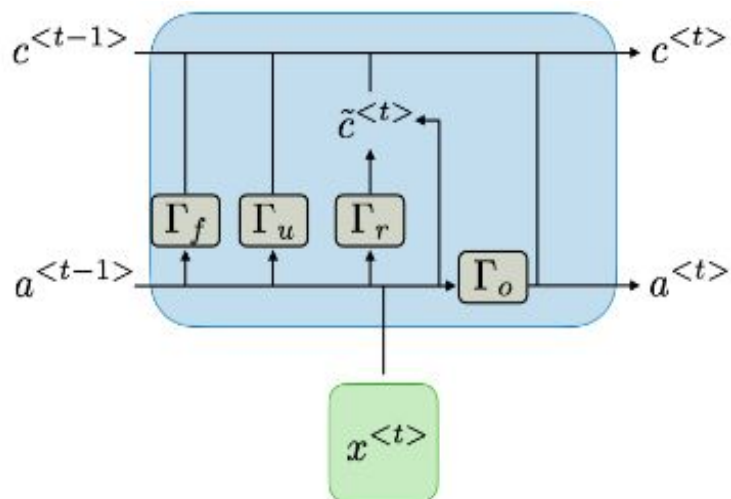
II . Kiến trúc hệ thống - mô hình

- LSTM: LSTM là một kiến trúc thuộc lớp mô hình RNN. Cấu tạo chung của LSTM là một kiến trúc có dạng truy hồi cho phép dự báo biến mục tiêu tuần tự theo thời gian. Các mô hình LSTM do đó thường được sử dụng phổ biến trong các tác vụ sequence-to-sequence như dịch máy, tóm tắt văn bản.



II . Kiến trúc hệ thống - mô hình

- GRU: là một dạng biến thể khác của LSTM, được gọi là cổng truy hồi đơn vị. Nó kết hợp cổng quên và cổng vào thành một cổng đơn giản gọi là cập nhật (update gate). Nó cũng nhập các ô trạng thái ẩn và thực hiện một số thay đổi khác. Kết quả của GRU đơn giản hơn nhiều so với LSTM.



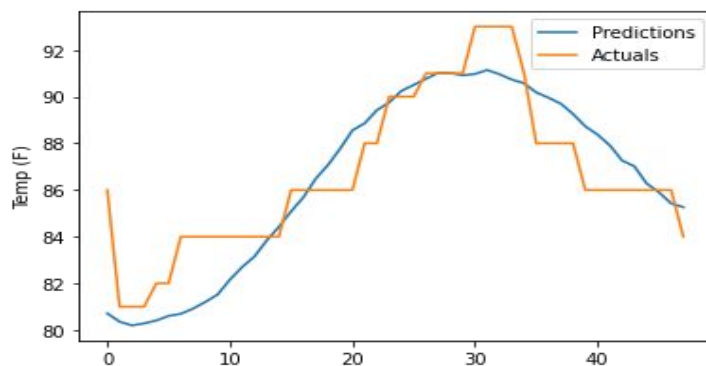


II . Kiến trúc hệ thống - mô hình

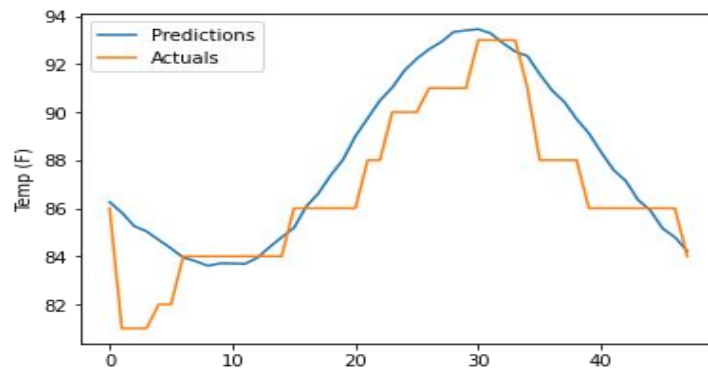
- Nhóm thực hiện huấn luyện và đánh giá các mô hình với các hyperparameters sau: epochs là 40, learning rate = 0.0001, hàm mục tiêu được lựa chọn là Mean squared error, metric là root mean square error, r2_score, mean absolute error và phương pháp optimizer là Adam.

II . Kiến trúc hệ thống - LSTM

- Sử dụng các trường: **Temperature, time (chuyển về day sin + day cos + year sin + year cos), humidity, wind speed, wind, pressure, condition.** Time step = 4

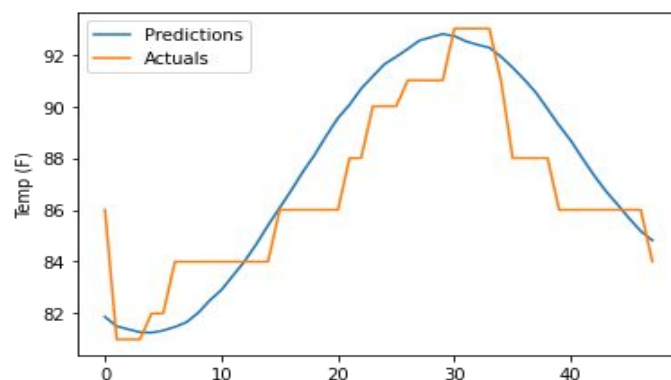


- Sử dụng trường **Temperature.** Time step = 4

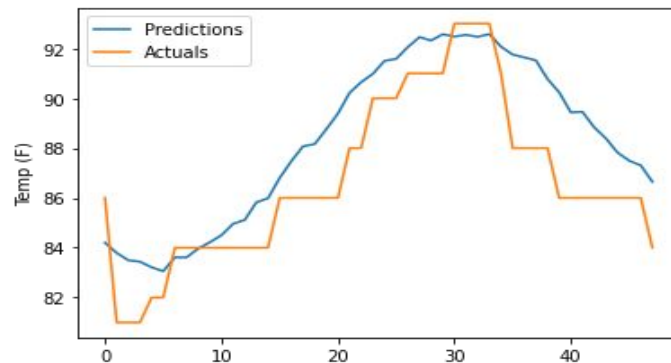


II . Kiến trúc hệ thống - GRU

- Sử dụng các trường: **Temperature, time (chuyển về day sin + day cos + year sin + year cos), humidity, wind speed, wind, pressure, condition.** Time step = 4



- Sử dụng trường **Temperature.** Time step = 4

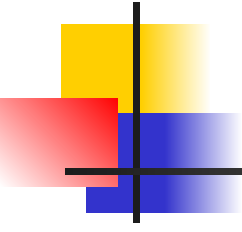




II . Kiến trúc hệ thống

- Kết quả trên tập test

Model	MSE	RMSE	MAE	R2_SCORE
LSTM + 10 trường thuộc tính + time step = 4	2.947	1.714	1.346	0.726
LSTM+ trường temperature + time step = 4	3.691	1.921	1.474	0.656
LSTM + 10 trường thuộc tính + time step = 1	3.197	1.788	1.508	0.820
GRU + 10 trường thuộc tính + time step = 4	3.255	1.804	1.487	0.697
GRU + trường temperature + time step = 4	4.397	2.097	1.802	0.590



DEMO
