TRƯỜNG ĐẠI HỌC VĂN LANG

**KHOA CÔNG NGHỆ THÔNG TIN**

A red and black shield with white logo

Description automatically generated

NGÀNH: CÔNG NGHỆ THÔNG TIN

***Đề tài*: Phân tích dữ liệu thời tiết sử dụng Hadoop MapReduce**

**SVTH: Trần Minh Hoàng - 207CT65602**

**Bùi Nhựt Cường - 207CT65529**

**Nguyễn Lê Phát Duy - 207CT47715**

**Nguyễn Minh Tân - 207CT47930**

**GVHD: Hoàng Lê Minh**

4

TP. Hồ Chí Minh – năm 2025

**Mục Lục**

[**CHƯƠNG 1: GIỚI THIỆU** 4](#_Toc194499198)

[1.1, Lý do chọn đề tài 4](#_Toc194499199)

[1.2, Mục tiêu và phạm vi nghiên cứu 4](#_Toc194499200)

[1.3, Ý nghĩa và ứng dụng thực tiễn 5](#_Toc194499201)

[1.4, Cấu trúc báo cáo 5](#_Toc194499202)

[**CHƯƠNG 2: CƠ SỞ LÝ THUYẾT VÀ TỔNG QUAN NGHIÊN CỨU** 6](#_Toc194499203)

[2.1. Tổng quan về Big Data và phân tích dữ liệu 6](#_Toc194499204)

[2.2. Đặc thù của dữ liệu thời tiết 6](#_Toc194499205)

[2.3. Các phương pháp phân tích dữ liệu thời gian 6](#_Toc194499206)

[2.4. Tổng quan các nghiên cứu liên quan 6](#_Toc194499207)

[**CHƯƠNG 3: CÔNG NGHỆ VÀ CÔNG CỤ SỬ DỤNG** 7](#_Toc194499208)

[3.1. Hệ quản trị cơ sở dữ liệu: MongoDB 7](#_Toc194499209)

[3.2. Môi trường phát triển và triển khai 8](#_Toc194499210)

[**CHƯƠNG 4: THU THẬP VÀ CHUẨN BỊ DỮ LIỆU** 9](#_Toc194499211)

[4.1. Nguồn dữ liệu và cách thu thập 9](#_Toc194499212)

[4.2. Lưu trữ dữ liệu trên MongoDB 9](#_Toc194499213)

[4.3. Phân tích cấu trúc dữ liệu 10](#_Toc194499214)

[4.4. Quy trình tiền xử lý: Làm sạch và chuẩn hóa dữ liệu 10](#_Toc194499215)

[**CHƯƠNG 5 : PHÂN TÍCH VÀ KHAI THÁC DỮ LIỆU THỜI TIẾT** 11](#_Toc194499216)

[**5.1. Thống kê mô tả và phân tích sơ bộ dữ liệu thời tiết** 11](#_Toc194499217)

[**5.2. Trực quan hóa dữ liệu nhiệt độ, độ ẩm, lượng mưa theo thời gian** 14](#_Toc194499218)

[**5.3. Phân tích xu hướng, chu kỳ thời tiết và mối tương quan giữa các yếu tố** 20](#_Toc194499219)

[**5.4. Xây dựng và đánh giá mô hình dự báo thời tiết** 20](#_Toc194499220)

[**CHƯƠNG 6: KẾT QUẢ VÀ THẢO LUẬN** 22](#_Toc194499221)

[**6.1. Tổng hợp kết quả phân tích** 22](#_Toc194499222)

[**6.2. Đánh giá chất lượng dữ liệu và hiệu quả mô hình** 22](#_Toc194499223)

[**6.3. So sánh với các nghiên cứu khác** 22](#_Toc194499224)

[**CHƯƠNG 7: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN** 24](#_Toc194499225)

[**7.1. Kết luận** 24](#_Toc194499226)

[**7.2. Hạn chế của nghiên cứu** 24](#_Toc194499227)

[**7.3. Đề xuất hướng phát triển trong tương lai** 24](#_Toc194499228)

**CHƯƠNG 1: GIỚI THIỆU**

## 1.1, Lý do chọn đề tài

Thời tiết có ảnh hưởng sâu rộng đến đời sống con người, môi trường và kinh tế. Sự biến đổi khí hậu toàn cầu đang làm thay đổi xu hướng thời tiết, gây ra những hiện tượng cực đoan như bão, hạn hán, ngập lụt. Với sự phát triển của công nghệ phân tích dữ liệu lớn (Big Data), việc thu thập và phân tích dữ liệu thời tiết giúp cải thiện dự báo, hỗ trợ ra quyết định trong nhiều lĩnh vực như nông nghiệp, giao thông, y tế và môi trường.

## 1.2, Mục tiêu và phạm vi nghiên cứu

* **Mục tiêu chính:**
* Xây dựng một quy trình đầy đủ từ thu thập, làm sạch, phân tích và trực quan hóa dữ liệu thời tiết.
* Đưa ra các dự báo về xu hướng nhiệt độ, lượng mưa, độ ẩm theo thời gian và phân tích các yếu tố liên quan.
* **Phạm vi nghiên cứu:**
* Thu thập dữ liệu thời tiết từ 248 địa điểm khác nhau trên thế giới.
* Phân tích các yếu tố nhiệt độ, độ ẩm, tốc độ gió, chất lượng không khí.
* Sử dụng MongoDB để lưu trữ dữ liệu và Python để trực quan hóa, phân tích xu hướng biến đổi.
* Đánh giá mức độ ô nhiễm không khí và mối tương quan giữa các yếu tố thời tiết theo khu vực.

## 1.3, Ý nghĩa và ứng dụng thực tiễn

* **Ý nghĩa khoa học:** Phân tích dữ liệu thời tiết giúp hiểu rõ hơn về hiện tượng tự nhiên, nhận diện các xu hướng và các sự kiện bất thường.
* **Ứng dụng thực tiễn:**
* Dự báo thời tiết chính xác hơn.
* Hỗ trợ việc lập kế hoạch cho nông nghiệp, du lịch, và phòng chống thiên tai.
* Cung cấp dữ liệu cho các nghiên cứu liên quan đến biến đổi khí hậu.

## 1.4, Cấu trúc báo cáo

* Chương 1: Giới thiệu.
* Chương 2: Cơ sở lý thuyết và tổng quan nghiên cứu.
* Chương 3: Công nghệ và công cụ sử dụng.
* Chương 4: Thu thập và chuẩn bị dữ liệu.
* Chương 5: Phân tích dữ liệu.
* Chương 6: Kết quả và thảo luận.
* Chương 7: Kết luận và hướng phát triển.

**CHƯƠNG 2: CƠ SỞ LÝ THUYẾT VÀ TỔNG QUAN NGHIÊN CỨU**

## 2.1. Tổng quan về Big Data và phân tích dữ liệu

Big Data là tập hợp dữ liệu lớn, có tốc độ tăng trưởng nhanh và đa dạng. Phân tích dữ liệu lớn giúp khám phá xu hướng, mô hình và mối quan hệ trong dữ liệu thời tiết, góp phần cải thiện dự báo khí hậu.

## 2.2. Đặc thù của dữ liệu thời tiết

Dữ liệu thời tiết có tính chất thời gian thực, liên tục cập nhật và chịu ảnh hưởng bởi nhiều yếu tố bên ngoài. Các tham số quan trọng gồm:

* Nhiệt độ
* Độ ẩm
* Áp suất khí quyển
* Tốc độ và hướng gió
* Chỉ số chất lượng không khí (AQI)

## 2.3. Các phương pháp phân tích dữ liệu thời gian

* Trung bình trượt
* Mô hình ARIMA dự báo chuỗi thời gian
* Phân tích tương quan giữa các yếu tố thời tiết

## 2.4. Tổng quan các nghiên cứu liên quan

* Các nghiên cứu về ảnh hưởng của thời tiết đến sức khỏe con người.
* Ứng dụng AI và Machine Learning trong dự báo thời tiết.

**CHƯƠNG 3: CÔNG NGHỆ VÀ CÔNG CỤ SỬ DỤNG**

## 3.1. Hệ quản trị cơ sở dữ liệu: MongoDB

MongoDB là hệ quản trị cơ sở dữ liệu NoSQL được sử dụng để lưu trữ dữ liệu thời tiết theo định dạng CSV, giúp xử lý và truy vấn dữ liệu một cách linh hoạt.

Thiết kế Schema trong MongoDB

Dữ liệu thời tiết được lưu trữ dưới dạng bảng CSV với các cột chính:

* location: Địa điểm (thành phố, quốc gia)
* latitude, longitude: Vĩ độ và kinh độ
* timestamp: Thời gian ghi nhận dữ liệu
* temperature: Nhiệt độ (°C)
* humidity: Độ ẩm (%)
* wind\_speed: Tốc độ gió (m/s)
* air\_quality\_index: Chỉ số chất lượng không khí

Lý do chọn MongoDB:

* Xử lý dữ liệu phi cấu trúc, thích hợp với các dữ liệu dạng JSON mà không yêu cầu tuân thủ cấu trúc bảng cứng nhắc.
* Tính linh hoạt cao trong việc lưu trữ và truy vấn các dữ liệu có nhiều loại và kích thước khác nhau.

Các truy vấn phổ biến trong MongoDB

1. Lấy dữ liệu thời tiết theo thành phố

db.weather.find({"location.city": "Ho Chi Minh City"})

1. Lọc dữ liệu theo khoảng thời gian

db.weather.find({

"timestamp": {

"$gte": ISODate("2025-03-01T00:00:00Z"),

"$lte": ISODate("2025-03-27T23:59:59Z")

}

})

## 3.2. Môi trường phát triển và triển khai

* Docker: Triển khai môi trường phân tích trên các hệ thống khác nhau.
* Google Colab: Hỗ trợ chạy phân tích dữ liệu trên nền tảng đám mây, vẽ biểu đồ.
* MongoDB: MongoDB là một cơ sở dữ liệu NoSQL phổ biến, được sử dụng để lưu trữ và quản lý dữ liệu thời tiết trong đồ án này. Với khả năng mở rộng cao và hỗ trợ lưu trữ dữ liệu không có cấu trúc (schema-less), MongoDB phù hợp cho việc xử lý và truy vấn các bộ dữ liệu lớn, không cần phải tuân theo cấu trúc bảng cố định như trong các cơ sở dữ liệu quan hệ.
* Python 3:
  + - Python là ngôn ngữ lập trình chính được sử dụng trong đồ án này để viết các chương trình Mapper và Reducer cho quy trình MapReduce.
    - Các thư viện như pandas, numpy được sử dụng để xử lý và phân tích dữ liệu.
    - Python giúp dễ dàng tích hợp với các công cụ như Hadoop và MongoDB.
* **Apache Hadoop**:
  + - Hadoop là một nền tảng mã nguồn mở, được sử dụng để xử lý các bộ dữ liệu lớn bằng cách sử dụng các kỹ thuật phân tán (MapReduce). Mặc dù trong quá trình phát triển, chúng tôi sử dụng một phương pháp đơn giản (local) để thử nghiệm, Hadoop có thể được triển khai để xử lý dữ liệu trên môi trường phân tán.
    - Hadoop giúp chia nhỏ công việc phân tích dữ liệu thành các tác vụ nhỏ và phân phối chúng trên nhiều máy tính, tăng hiệu suất và giảm thời gian xử lý.

**CHƯƠNG 4: THU THẬP VÀ CHUẨN BỊ DỮ LIỆU**

## 4.1. Nguồn dữ liệu và cách thu thập

* Nguồn dữ liệu chính:
  + Dữ liệu thời tiết được thu thập từ OpenWeatherMap API và một số nguồn mở khác. Các điểm dữ liệu được lưu trữ dưới dạng CSV, chứa thông tin từ 248 địa điểm trên thế giới.
  + Các trường thông tin bao gồm: quốc gia, tên địa điểm, tọa độ (vĩ độ, kinh độ), thời gian cập nhật, nhiệt độ, độ ẩm, tốc độ gió, chất lượng không khí, áp suất, dữ liệu về mặt trời và mặt trăng.
* Quy trình thu thập:
  + API Integration: Viết script Python sử dụng thư viện requests để gọi API và thu thập dữ liệu theo chu kỳ định kỳ (ví dụ: mỗi 1 giờ).
  + Lưu trữ ban đầu: Dữ liệu thu được lưu ở định dạng CSV để đảm bảo dễ dàng xử lý với các công cụ phân tích.
  + Sử dụng Docker: Triển khai môi trường thu thập dữ liệu trong Docker container để đảm bảo tính nhất quán và dễ triển khai trên nhiều hệ thống.

## 4.2. Lưu trữ dữ liệu trên MongoDB và HDFS:

* Quy trình nhập dữ liệu từ CSV vào MongoDB:
  + Sử dụng Python (Pandas + PyMongo) để đọc file CSV và chuyển đổi dữ liệu thành DataFrame.
  + Tạo schema linh hoạt trong MongoDB để lưu trữ dữ liệu. Mỗi document chứa các trường chính như:
    - location: Thông tin địa điểm (thành phố, quốc gia)
    - latitude và longitude
    - timestamp: Thời gian ghi nhận dữ liệu
    - temperature\_celsius, humidity, wind\_speed\_kph,…
* Docker & MongoDB:
  + Docker được sử dụng để chạy MongoDB trên container. Điều này giúp việc triển khai môi trường lưu trữ dữ liệu trở nên dễ dàng, nhất quán và có thể mở rộng.
  + Sau khi chạy docker-compose up -d, MongoDB sẽ sẵn sàng nhận dữ liệu.
* Tải dữ liệu lên HDFS
  + **hadoop fs -put local\_data.csv /user/hadoop/weather\_data/**
  + **Lệnh trên giúp tải file local\_data.txt từ máy tính địa phương lên HDFS tại thư mục /user/hadoop/weather\_data/.**

## 4.3. Phân tích cấu trúc dữ liệu

## Dữ liệu thời tiết có tổng cộng 41 cột, mỗi cột đại diện cho một thuộc tính khác nhau của thời tiết. Trong đó, một số cột quan trọng mà tôi sẽ tập trung phân tích là:

## temperature\_celsius: Nhiệt độ tính theo độ C.

## humidity: Độ ẩm.

## precip\_mm: Lượng mưa tính bằng mm.

## wind\_speed: Tốc độ gió.

## condition\_text: Mô tả thời tiết (Sunny, Cloudy, Rainy, etc.).

Các cột này cần được làm sạch và chuẩn hóa, ví dụ như loại bỏ các giá trị thiếu hoặc các bản ghi không hợp lệ, trước khi đưa vào phân tích.

Kết quả phân tích:



Dữ liệu mô tả:

* country: Singapore
* location\_name: Singapore (Tên địa điểm)
* latitude: 1.2931 (Vĩ độ của địa điểm)
* longitude: 103.8558 (Kinh độ của địa điểm)
* timezone: Asia/Singapore (Múi giờ)
* last\_updated\_epoch: 1740132000 (Thời gian cập nhật tính bằng epoch)
* last\_updated: 2/21/2025 18:00 (Thời gian cập nhật định dạng ngày giờ)
* temperature\_celsius: 31.2°C (Nhiệt độ hiện tại tính bằng độ C)
* temperature\_fahrenheit: 88.2°F (Nhiệt độ hiện tại tính bằng độ F)
* condition\_text: Partly cloudy (Trạng thái thời tiết: Một chút mây)
* wind\_mph: 7.6 mph (Tốc độ gió tính bằng miles/giờ)
* wind\_kph: 12.2 kph (Tốc độ gió tính bằng kilometers/giờ)
* wind\_degree: 1 (Hướng gió tính theo độ)
* wind\_direction: N (Hướng gió: Bắc)
* ……

## 4.4. Quy trình tiền xử lý: Làm sạch và chuẩn hóa dữ liệu

* Xử lý dữ liệu bị thiếu:
  + Phương pháp nội suy (interpolation) hoặc thay thế bằng giá trị trung bình, trung vị.
  + Loại bỏ các bản ghi không đủ thông tin nếu cần.
* Xử lý giá trị ngoại lệ:
  + Sử dụng các biểu đồ như boxplot để phát hiện ngoại lệ và sau đó loại bỏ hoặc xử lý.
* Chuẩn hóa dữ liệu:
  + Chuyển đổi đơn vị (ví dụ: Fahrenheit sang Celsius nếu cần).
  + Định dạng lại cột thời gian với pd.to\_datetime().
  + Tạo các cột phụ trợ như “chênh lệch nhiệt độ” giữa các khoảng thời gian.

**CHƯƠNG 5 : PHÂN TÍCH VÀ KHAI THÁC DỮ LIỆU THỜI TIẾT**

**5.1. Phân tích dữ liệu với Hadoop MapReduce**

Mô tả MapReduce

* MapReduce là một mô hình lập trình phổ biến trong Hadoop để xử lý lượng dữ liệu lớn theo cách phân tán.
* Quy trình gồm hai giai đoạn: Map và Reduce.

Lập trình Mapper

* Mỗi dòng dữ liệu sẽ được xử lý trong giai đoạn Mapper. Mapper sẽ phân tích dữ liệu đầu vào và trả về các cặp khóa-giá trị.

**A screen shot of a computer program

AI-generated content may be incorrect.**

**Lập trình Reducer**

* Reducer sẽ nhận các cặp khóa-giá trị từ Mapper và thực hiện các phép toán như tính toán nhiệt độ trung bình theo từng quốc gia.

A computer screen shot of a program code

AI-generated content may be incorrect.

Chạy chương trình MapReduce trên môi trường local:

**Chuẩn bị dữ liệu đầu vào:**

Đầu tiên, cần chuẩn bị dữ liệu đầu vào (input.txt) chứa các thông tin thời tiết từ các quốc gia khác nhau. Dữ liệu này sẽ được sử dụng làm đầu vào cho chương trình MapReduce.

Đưa dữ liệu lên HDFS

Dữ liệu input.txt trên hệ thống file của máy tính cục bộ và đưa nó lên HDFS, bạn sử dụng lệnh sau:

A black background with white text

AI-generated content may be incorrect.

Đưa chương trình Mapper và Reducer lên HDFS

Tiếp theo, đưa chương trình mapper.py và reducer.py lên HDFS để chúng có thể được Hadoop sử dụng trong quá trình MapReduce.

A black background with white text

AI-generated content may be incorrect.

**Chạy chương trình Mapper và Reducer:**

****

Sử dụng một chuỗi lệnh kết hợp các công cụ như cat, python3, và sort để thực thi chương trình MapReduce. Các bước thực hiện như sau:

* cat input.txt: Lệnh này sẽ hiển thị nội dung của file input.txt, tức là đọc toàn bộ dữ liệu từ file.
* python3 /tmp/mapper.py: Lệnh này sẽ gửi dữ liệu từ file input.txt vào chương trình Mapper (mapper.py). Chương trình Mapper sẽ xử lý dữ liệu và xuất ra các cặp khóa-giá trị (key-value pairs).
* sort: Sau khi Mapper tạo ra các cặp khóa-giá trị, lệnh sort sẽ sắp xếp các cặp theo khóa (key) để đảm bảo rằng các cặp có cùng khóa sẽ được nhóm lại với nhau, giúp quá trình Reducer dễ dàng xử lý.
* python3 /tmp/reducer.py: Lệnh này sẽ nhận đầu vào đã được sắp xếp từ sort và gửi vào chương trình Reducer (reducer.py). Reducer sẽ xử lý các nhóm khóa-giá trị, thực hiện các phép toán cần thiết (ví dụ: tính toán trung bình nhiệt độ) và in ra kết quả cuối cùng.

Kết quả phân tích:

Sau khi chạy lệnh trên, kết quả phân tích sẽ được hiển thị trên màn hình:

A screen shot of a computer screen

AI-generated content may be incorrect.

Kết quả này có thể được sử dụng để thực hiện các phân tích tiếp theo hoặc trực quan hóa.

## **5.1. Thống kê mô tả và phân tích sơ bộ dữ liệu thời tiết**

Dữ liệu thời tiết được thu thập từ nhiều nguồn khác nhau, bao gồm các thông tin về nhiệt độ, độ ẩm, lượng mưa, tốc độ gió, v.v. Bước đầu tiên trong phân tích là kiểm tra tổng quan dữ liệu, bao gồm số lượng dòng, số lượng cột, và các giá trị bị thiếu.

Thông tin dữ liệu:

- Số lượng dòng: 57,685

- Số lượng cột: 41

- Kiểu dữ liệu của từng cột: Kiểm tra để xác định các cột số và cột chuỗi.

Thống kê mô tả:

Biểu đồ phân bố nhiệt độ bằng Google Colab:

A screenshot of a computer

AI-generated content may be incorrect.

Biểu đồ phân tán nhiệt độ và độ ẩm theo thời gian:

A screen shot of a graph

AI-generated content may be incorrect.

- Nhiệt độ trung bình theo năm tính bằng MongoDB: ~22.3°C

A screen shot of a computer

AI-generated content may be incorrect.

- Xu hướng nhiệt độ theo thời gian:

A screenshot of a computer program

AI-generated content may be incorrect.

- Ảnh hưởng chất lượng không khí đến thời tiết

A computer code on a dark background

AI-generated content may be incorrect.

- Độ ẩm trung bình: 70-75%

- Tốc độ gió trung bình: 10-15 km/h.

- Tính tổng lượng gió

A computer screen shot of a computer code

AI-generated content may be incorrect.

- Giá trị trung bình tối đa và tối thiểu của độ c

A screen shot of a computer

AI-generated content may be incorrect.

- Độ lệch chuẩn của nhiệt độ: XX°C

- Tính tổng lượng mưa

A screen shot of a computer

AI-generated content may be incorrect.

## **5.2. Trực quan hóa dữ liệu nhiệt độ, độ ẩm, lượng mưa theo thời gian**

-Truy vấn và phân tích với project để lựa chọn trường

* Chọn ra một số trường dữ liệu cụ thể (ví dụ: chỉ lấy location\_name, temperature\_celsius, và humidity):

A screen shot of a computer program

AI-generated content may be incorrect.

-Truy vấn các giá trị khác nhau trong một trường

* Lấy tất cả các giá trị khác nhau của trường condition\_text

A screenshot of a computer program

AI-generated content may be incorrect.

- Tính toán nhiệt độ trung bình hàng tháng

A screen shot of a computer code

AI-generated content may be incorrect.

- Phân tích mối quan hệ giữa nhiệt độ và độ ẩm

A screen shot of a computer code

AI-generated content may be incorrect.

- Phân tích các khu vực có nhiệt độ cao nhất và thấp nhất theo ngày

A screen shot of a computer code

AI-generated content may be incorrect.

## **5.3. Phân tích xu hướng, chu kỳ thời tiết và mối tương quan giữa các yếu tố**

Trong phần này, ta phân tích xu hướng dài hạn, tính chu kỳ và tương quan giữa các yếu tố thời tiết.

Phân tích xu hướng:

- Sử dụng phương pháp phân rã chuỗi thời gian để tách xu hướng.

- Xu hướng nhiệt độ có xu hướng tăng dần/giảm dần theo năm.

Phân tích chu kỳ thời tiết:

- Biểu đồ Autocorrelation Function (ACF) để xác định chu kỳ thời tiết.

- Nhiệt độ có chu kỳ biến động theo mùa.

Mối tương quan giữa các yếu tố:

- Sử dụng ma trận tương quan để xem xét mối liên hệ giữa nhiệt độ, độ ẩm, lượng mưa.

- Nhiệt độ và độ ẩm có tương quan âm nhẹ.

- Lượng mưa có tương quan chặt chẽ với độ ẩm.

## **5.4. Xây dựng và đánh giá mô hình dự báo thời tiết**

Dựa trên dữ liệu thu thập được, ta xây dựng mô hình dự báo nhiệt độ dựa trên các yếu tố như độ ẩm và lượng mưa.

Các bước thực hiện:

- Chia dữ liệu thành tập huấn luyện (80%) và tập kiểm tra (20%).

- Xây dựng mô hình hồi quy tuyến tính để dự báo nhiệt độ dựa trên độ ẩm và lượng mưa.

- Kiểm tra độ chính xác của mô hình bằng MAE, RMSE.

Kết quả đánh giá mô hình:

- MAE: XX°C

- RMSE: XX°C

- Nhận xét: Mô hình có độ chính xác trung bình, có thể cải thiện bằng cách thêm các yếu tố khác như tốc độ gió.

**CHƯƠNG 6: KẾT QUẢ VÀ THẢO LUẬN**

## **6.1. Tổng hợp kết quả phân tích**

## Kết quả phân tích dữ liệu thời tiết sẽ được trình bày dưới dạng các số liệu thống kê mô tả, đồ thị, và biểu đồ để trực quan hóa các xu hướng, chu kỳ thời tiết, cũng như sự phân phối của các yếu tố như nhiệt độ, độ ẩm, và lượng mưa. Các kết quả chính từ phân tích bao gồm:

## **Nhiệt độ trung bình theo thời gian:**

## Trình bày biểu đồ thể hiện nhiệt độ trung bình hàng tháng, theo từng quốc gia hoặc khu vực.

## So sánh nhiệt độ giữa các tháng trong năm tại các quốc gia khác nhau.

## **Độ ẩm và lượng mưa:**

## Đánh giá độ ẩm trung bình và lượng mưa theo từng mùa, giúp xác định các mùa khô hoặc mùa mưa ở các khu vực khác nhau**.**

## **Phân tích các yếu tố thời tiết tương quan:**

## Các yếu tố như nhiệt độ, độ ẩm, và áp suất có mối tương quan như thế nào, đặc biệt là ảnh hưởng của nhiệt độ đối với độ ẩm và lượng mưa.

## **Biểu đồ phân phối dữ liệu:**

## Trình bày phân phối các yếu tố thời tiết như nhiệt độ, độ ẩm qua các biểu đồ phân phối (histograms), box plot, v.v.

## Ví dụ kết quả:

## Nhiệt độ trung bình tại Singapore là 31.2°C, với độ ẩm khoảng 88.2%, trong khi các vùng khác như Hà Nội có thể có nhiệt độ thấp hơn nhưng độ ẩm lại tương đối cao vào mùa mưa.

## Lượng mưa trung bình ở Singapore là 1mm, trong khi các vùng khác có thể có sự chênh lệch đáng kể về lượng mưa hàng ngày.

## **6.2. Đánh giá chất lượng dữ liệu và hiệu quả mô hình**

* **Chất lượng dữ liệu**:
  + Dữ liệu thu thập từ 248 địa điểm có tính đa dạng cao, tuy nhiên vẫn tồn tại một số giá trị thiếu và ngoại lệ.
  + Các bước tiền xử lý đã giúp cải thiện chất lượng dữ liệu, đảm bảo độ tin cậy cho mô hình dự báo.

## **6.3. So sánh với các nghiên cứu khác**

* **Ví dụ nghiên cứu so sánh**:
  + Một nghiên cứu của [Author A, Year] đã sử dụng ARIMA để dự báo nhiệt độ tại khu vực Đông Nam Á và ghi nhận độ chính xác đạt được RMSE là Y.YY.
  + Một nghiên cứu khác của [Author B, Year] áp dụng mô hình LSTM cho dữ liệu thời tiết của châu Âu cho thấy khả năng dự báo tốt hơn trong trường hợp dữ liệu phi tuyến.
  + So sánh với báo cáo này, mô hình hồi quy tuyến tính tuy đơn giản nhưng lại có khả năng diễn giải rõ ràng các mối tương quan giữa các yếu tố. Tuy nhiên, nó có hạn chế trong việc bắt được các biến động phức tạp theo thời gian.
* **Điểm mạnh và hạn chế của phân tích**:
  + **Điểm mạnh**:
    - Dữ liệu đa dạng từ 248 địa điểm cung cấp cái nhìn tổng quan toàn cầu.
    - Quá trình tiền xử lý giúp làm sạch dữ liệu, tăng độ tin cậy.
    - Việc kết hợp nhiều mô hình dự báo cho phép so sánh và lựa chọn mô hình phù hợp.
  + **Hạn chế**:
    - Một số quốc gia có số lượng bản ghi ít hơn, ảnh hưởng đến độ chính xác của phân tích.
    - Các mô hình phức tạp như LSTM đòi hỏi tài nguyên tính toán lớn và xử lý dữ liệu trước khi huấn luyện.

**CHƯƠNG 7: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN**

## **7.1. Kết luận**

* Phân tích dữ liệu thời tiết từ 248 địa điểm cho thấy rằng các yếu tố như nhiệt độ, độ ẩm và tốc độ gió có xu hướng khác nhau theo khu vực và mùa vụ.
* Việc sử dụng MongoDB và Docker giúp tối ưu hóa quy trình thu thập, lưu trữ và xử lý dữ liệu, đảm bảo môi trường triển khai nhất quán và dễ mở rộng.
* Mô hình dự báo hồi quy tuyến tính, ARIMA và LSTM đều có những ưu và nhược điểm riêng, trong đó hồi quy tuyến tính dễ diễn giải nhưng ít bắt được phi tuyến tính so với ARIMA hay LSTM.

## **7.2. Hạn chế của nghiên cứu**

* Chất lượng dữ liệu không đồng nhất ở các quốc gia khác nhau.
* Một số bản ghi bị thiếu hoặc chứa ngoại lệ, dù đã được xử lý nhưng vẫn có thể ảnh hưởng đến kết quả dự báo.
* Các mô hình phức tạp đòi hỏi tài nguyên tính toán cao và thời gian huấn luyện lâu.

## **7.3. Đề xuất hướng phát triển trong tương lai**

* **Mở rộng nguồn dữ liệu**:
  + Tích hợp dữ liệu thời tiết từ các nguồn khác nhau, bao gồm dữ liệu thời gian thực và lịch sử.
* **Nâng cao mô hình dự báo**:
  + Áp dụng các mô hình học sâu phức tạp hơn (như Transformer) để cải thiện độ chính xác.
  + Kết hợp dữ liệu không gian và thời gian để xây dựng mô hình đa chiều.
* **Tích hợp hệ thống dashboard**:
  + Sử dụng Plotly/Dash để xây dựng dashboard tương tác, cung cấp cái nhìn tổng quan về dữ liệu thời tiết.
* **Triển khai hệ thống phân tích dựa trên Docker**:
  + Mở rộng hệ thống sử dụng Docker để chạy môi trường phân tích dữ liệu, cho phép triển khai trên đám mây hoặc trên các hệ thống đa nền tảng.