ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KHOA HỌC VÀ KĨ THUẬT MÁY TÍNH

BÁO CÁO
ĐỒ ÁN CHUYÊN NGÀNH

XÂY DỰNG CƠ SỞ DỮ LIỆU TÍCH HỢP CHO CÁC HỆ THỐNG DỰ BÁO NGẮN HẠN TRONG KHÍ TƯỢNG THUỶ VĂN

Ngành: Khoa học máy tính

HỘI ĐỒNG: **Hội đồng 5**
GVHD1: **Lê Hồng Trang**
GVHD2: **Trương Quỳnh Chi**
TKHĐ: **Lê Thị Bảo Thu**
—o0o—
SVTH1: **Trần Hà Tuấn Kiệt (2011493)**
SVTH2: **Nguyễn Đức Thuỵ (2012158)**

Thành phố Hồ Chí Minh, Tháng 12/2023

Lời cam đoan

Chúng tôi cam đoan rằng đồ án này là kết quả của công việc nghiên cứu của chúng tôi, được thực hiện dưới sự hướng dẫn tận tâm của PGS. TS. Lê Hồng Trang tại Trường Đại Học Bách Khoa - Đại học Quốc gia TP.HCM . Chúng tôi xác nhận rằng tất cả các thành viên trong nhóm đã đóng góp công sức và kiến thức của mình để hoàn thành công trình này.

Chúng tôi cam đoan rằng đồ án này không phải là bản sao từ bất kỳ công trình nào khác và tất cả các nguồn thông tin đã được chú thích đầy đủ theo quy tắc trích dẫn khoa học. Mọi ý kiến, thông tin hay dữ liệu từ nguồn ngoại lai đều được công bố và thực hiện theo quy tắc đạo đức và quy định của Trường Đại Học Bách Khoa - Đại học Quốc gia TP.HCM .

Chúng tôi xác nhận rằng tất cả các thành viên trong nhóm đã thực hiện nghiên cứu này với lòng tận tụy và tính chân thành cao, tuân thủ mọi nguyên tắc đạo đức nghiên cứu khoa học. Chúng tôi cũng đảm bảo rằng không có hành vi gian lận hoặc vi phạm nào đã xảy ra trong quá trình thực hiện đồ án.

Chúng tôi chấp nhận mọi trách nhiệm và hậu quả pháp lý liên quan đến nội dung của đồ án và đồng thời biết ơn sự hướng dẫn, góp ý quý báu của giáo viên PGS. TS. Lê Hồng Trang đã giúp chúng tôi hoàn thiện công trình này.

Thành phố Hồ Chí Minh, Ngày 29 tháng 12 năm 2023

NGƯỜI THỰC HIỆN

Trần Hà Tuấn Kiệt

Nguyễn Đức Thuỵ

Lời cảm ơn / Lời ngỏ

Chúng tôi, nhóm nghiên cứu gồm hai thành viên, xin gửi lời cảm ơn chân thành đến tất cả mọi người đã đóng góp vào thành công của đồ án này.

Đầu tiên và quan trọng nhất, chúng tôi muốn bày tỏ lòng biết ơn đến PGS. TS. Lê Hồng Trang . Sự tận tâm và kiến thức sâu rộng của thầy không chỉ hướng dẫn chúng tôi qua những thách thức của đồ án mà còn giúp chúng tôi phát triển kỹ năng nghiên cứu và phê bình.

Chúng tôi cũng muốn bày tỏ lòng biết ơn đặc biệt đến tất cả những người bạn, đồng nghiệp, và gia đình đã hỗ trợ chúng tôi trong suốt quá trình nghiên cứu. Sự góp ý và ý kiến của mọi người đã làm giàu thêm nội dung và chất lượng của đồ án.

Cuối cùng, nhưng không kém phần quan trọng, chúng tôi cảm ơn nhau - đối tác nghiên cứu đồng hành trong mỗi bước của đồ án. Sự hợp tác và đóng góp chung của chúng tôi đã tạo nên một sản phẩm mà chúng tôi tự hào.

Chúng tôi tin rằng đồ án này là một bước tiến quan trọng trong sự phát triển của chúng tôi và không thể đạt được mà không có sự hỗ trợ và đóng góp của tất cả mọi người.

Tóm tắt nội dung

Trong tài liệu này, chúng tôi giới thiệu một đề xuất toàn diện để tích hợp, điều phối và giám sát dữ liệu khí tượng thủy văn tại Việt Nam. Hệ thống được đề xuất có thể làm nền tảng cho việc xây dựng một kho dữ liệu quốc gia chuyên sâu về thông tin khí tượng.

Để làm rõ đề xuất của chúng tôi, chúng tôi đã xây dựng một bản thử nghiệm toàn diện, trong đó chúng tôi đã thực hiện mô phỏng về quy trình thu thập và truyền phát dữ liệu từ trạm thời tiết tại Nhà Bè. Đồng thời, chúng tôi đã tập trung vào quá trình chuyển đổi và tổ chức lưu trữ dữ liệu để đảm bảo sự minh bạch và hiệu suất tối ưu trong quá trình xử lý thông tin.

Bài nghiên cứu này là biểu hiện của những nỗ lực đáng kể trong việc tối ưu hóa quản lý dữ liệu khí tượng tại Việt Nam. Hệ thống đề xuất này không chỉ giải quyết các thách thức về tích hợp và điều phối mà còn đặt nền tảng cho một kho dữ liệu quốc gia vững mạnh, phục vụ cho những nhu cầu đa dạng của nghiên cứu và ứng dụng trong lĩnh vực khí tượng.

Bảng phân công công việc

STT	Họ và tên	MSSV	Phân chia công việc
1	Trần Hà Tuấn Kiệt	2011493	- Phân tích dữ liệu - Xây dựng kiến trúc hệ thống - Nghiên cứu, tìm hiểu công nghệ và áp dụng - Xây dựng API
2	Nguyễn Đức Thuỵ	2012158	- Phân tích dữ liệu - Xây dựng kiến trúc hệ thống - Nghiên cứu, tìm hiểu công nghệ và áp dụng - Xây dựng Data Flow

Mục lục

Danh sách hình vẽ

Danh sách bảng

Chương 1

Giới thiệu

1.1 Phát biểu bài toán

1.2 Động lực thực hiện

Đối mặt với sự phức tạp ngày càng tăng của biến đổi khí hậu, các mô hình dự báo cũng không ngừng đòi hỏi tính chính xác ngày càng cao. Việc tích hợp thông tin từ nhiều nguồn và lưu trữ chúng trong một cơ sở dữ liệu hiệu quả trở thành chìa khóa quan trọng để đáp ứng mọi thách thức trong lĩnh vực này. Nhận thức được tầm quan trọng của dự báo thời tiết và thuỷ văn trong việc đảm bảo an toàn và phát triển cộng đồng, chúng tôi đề xuất dự án "Xây dựng cơ sở dữ liệu tích hợp cho các hệ thống dự báo ngắn hạn trong khí tượng thuỷ văn".

1.3 Mục tiêu dự án

Mục tiêu của dự án là nghiên cứu và xây dựng một cơ sở dữ liệu linh hoạt, có khả năng tích hợp dữ liệu, tổng hợp thông tin từ nhiều nguồn ở các trạm ra-đa và quan trắc khí tượng thuỷ văn, đồng thời lưu trữ chúng một cách hiệu quả. Dự án nhằm tạo ra một nguồn thông tin đáng tin cậy và toàn diện để nâng cao độ chính xác của mô hình dự báo thời tiết. Bằng cách này, chúng ta có thể hỗ trợ hiệu quả trong việc dự đoán các biến động thời tiết ngắn hạn. Cơ sở dữ liệu sẽ không chỉ linh hoạt mà còn hiệu quả, có khả năng tổng hợp thông tin một cách hiệu quả và lưu trữ chúng một cách có tổ chức. Qua đó, dự án không chỉ mang lại nguồn thông tin đáng tin cậy mà còn đóng góp vào việc hiểu rõ hơn về biến động thời tiết, đặc biệt là trong bối cảnh biến đổi khí hậu ngày càng trở nên quan trọng.

1.4 Phạm vi dự án

Dự án sẽ tập trung vào Thành phố Hồ Chí Minh, một đô thị lớn với môi trường khí hậu đặc biệt và có ảnh hưởng lớn đến cuộc sống hàng ngày của cộng đồng.

Chúng tôi sẽ nghiên cứu và thu thập dữ liệu từ nhiều trạm ra-đa và trạm quan trắc khí tượng thuỷ văn trong thành phố để đảm bảo tính đa dạng và đại diện cho các điều kiện thời tiết địa phương. Sau khi đã thu thập đủ dữ liệu cần thiết, chúng tôi sẽ tiến hành thực hiện các bước tiền xử lý và chuẩn hoá dữ liệu để đảm bảo độ chính xác và tính nhất quán của tập dữ liệu. Cuối cùng, chúng tôi sẽ xây dựng cơ sở dữ liệu tích hợp để lưu trữ và quản lý thông tin từ các nguồn

khác nhau, tạo ra một nguồn dữ liệu đồng nhất và đáng tin cậy. Nguồn thông tin tích hợp sẽ được sử dụng để cung cấp dữ liệu đa chiều và chi tiết về điều kiện thời tiết hiện tại và dự đoán ngắn hạn. Mục tiêu là giúp cộng đồng và các đơn vị liên quan chuẩn bị tốt hơn cho biến động thời tiết, đặc biệt là những thay đổi không dự đoán được.

Chúng tôi mong đợi rằng việc có một cơ sở dữ liệu tích hợp sẽ cải thiện hiệu quả của các mô hình dự báo thời tiết, giúp người dân và doanh nghiệp đối mặt với thời tiết khó khăn một cách thông minh và an toàn.

Chương 2

Cơ sở lý thuyết

2.1 Hệ Cơ sở dữ liệu

Database systems perform vital functions for all sorts of organizations because of the growing importance of using and managing data efficiently. A database system consists of a software, a database management system (DBMS) and one or several databases. DBMS is a set of programs that enables users to store, manage and access data. In other words database is processed by DBMS, which runs in the main memory and is controlled by the respective operating system

A database is a logically coherent collection of data with some inherent meaning and represents some aspects of the real world. A random assortment of data cannot be referred to as a database. Databases draw a sharp distinction between data and information. Data are known facts that can be recorded and that have implicit meaning. Information is data that have been organized and prepared in a form that is suitable for decision-making. Shortly information is the analysis and synthesis of data. The most fundamental terms used in database approach are ìentityî, ìattributeî and ìrelationshipî. An entity is something that can be identified in the usersí work environment, something that the users want to track. It may be an object with a physical or conceptual existence. An attribute is a property of an entity. A particular entity will have a value for each of its attributes. The attribute values that describe each entity become a major part of data stored in the database

Database Management System is a general-purpose software system designed to manage large bodies of information facilitating the process of defining, constructing and manipulating databases for various applications. Specifying data types, structures and constraints for the data to be stored in the database is called defining a database. Constructing the database is the process of storing data itself on some storage medium that is controlled by the DBMS. Querying to retrieve specific data, updating the database to reflect changes and generating reports from the data are the main concepts of manipulating a database. The DBMS functions as an interface between the users and the database ensuring that the data is stored persistently over long periods of time, independent of the programs that access it [3]. DBMS can be divided into three subsystems; the design tools subsystem, the run time subsystem and the DBMS engine.

The design tools subsystem has a set of tools to facilitate the design and creation of the database and its applications. Tools for creating tables, forms, queries and reports are components of this system. DBMS products also provide programming languages and interfaces to programming languages. The run time subsystem processes the application components that are developed using the design tools. The last component of DBMS is the DBMS engine which receives requests from the other two components and translates those requests into commands

to the operating system to read and write data on physical media. [4]

Database approach has several advantages over traditional file processing in which each user has to create and define files needed for a specific application. In these systems duplication of data is generally inevitable causing wasted storage space and redundant efforts to maintain common data up-to date. In database approach data is maintained in a single storage medium and accessed by various users. The self-describing nature of database systems provides information not only about database itself but also about the database structure such as the type and format of the data. A complete definition and description of database structure and constraints, called meta-data, is stored in the system catalog. Data abstraction is a consequence of this self-describing nature of database systems allowing programdata independence. DBMS access programs do not require changes when the structure of the data files are changed hence the description of data is not embedded in the access programs. This property is called program-data independence. Support of multiple views of data is another important feature of database systems, which enables different users to view different perspective of database dependent on their requirements. In a multi-user database environment users probably have access to the same data at the same time as well as they can access different portions of database for modification. Concurrency control is crucial for a DBMS so that the results of the updates are correct. The DBMS software is to ensure that concurrent transactions operate correctly when several users are trying to update the same data

Using a DBMS also eliminates unnecessary data redundancy. In database approach each primary fact is generally recorded in only one place in the database [6]. Sometimes it is desirable to include some limited redundancy to improve the performance of queries when it is more efficient to retrieve data from a single file instead of searching and collecting data from several files, but this data duplication is controlled by DBMS so as to prohibit inconsistencies among files. By eliminating data redundancy inconsistencies among data are also reduced [5]. Reducing reduncancy improves the consistency of data while reducing the waste in storage space. DBMS gives the opportunity of data sharing to the users. Sharing data often permits new data processing applications to be developed without having to create new data files. In general, less redundancy and greater sharing lead to less confusion between organizational units and less time spent resolving errors and inconsistencies in reports. The database approach also permits security restrictions. In a DBMS different types of authorizations are accepted in order to regulate which parts of the database various users can access or update.

2.2 Hệ thống dữ liệu

2.2.1 Data Flow

Dòng Dữ liệu (Data Flow) là sự chuyển động của dữ liệu từ một vị trí đến vị trí khác hoặc từ một quy trình này đến một quy trình khác trong hệ thống [3]. Data Flow thường có nơi dữ liệu bắt nguồn và nơi nó được tiêu thụ hoặc lưu trữ. Trong quá trình Data Flow được thực thi, dữ liệu có thể xảy ra các thay đổi thành định dạng hoặc cấu trúc khác.

Chuyển đổi dữ liệu (Data Transformation):

Chuyển đổi dữ liệu là một phần quan trọng của quá trình dòng dữ liệu, nơi dữ liệu được thay đổi để đáp ứng yêu cầu cụ thể của quy trình hoặc hệ thống. Có hai hướng chính cho việc thực hiện chuyển đổi dữ liệu: theo lô và theo thời gian thực.

Trong chế độ xử lý theo lô, dữ liệu được xử lý theo từng đợt, thường được lên lịch để xử lý vào các khoảng thời gian đặt trước. Điều này thích hợp cho các tác vụ yêu cầu xử lý dữ liệu lớn và phức tạp mà không cần đáp ứng ngay lập tức.

Ngược lại, xử lý theo thời gian thực là quá trình xử lý dữ liệu ngay khi nó đến, mà không có đợi đến khi có một lượng lớn dữ liệu để xử lý. Điều này thường được ưa chuộng trong các ứng dụng đòi hỏi độ trễ thấp, như xử lý sự kiện thời gian thực.

ETL:

ETL là một phương pháp quan trọng được sử dụng cho việc quản lý dòng dữ liệu trong hệ thống lưu trữ dữ liệu. Viết tắt ETL đến từ ba bước chính trong quy trình này.

1. **Extract (Trích Xuất):** Dữ liệu được trích xuất từ các nguồn khác nhau, chẳng hạn như cơ sở dữ liệu, tệp tin, hoặc các dịch vụ trực tuyến.
2. **Transform (Biến Đổi):** Dữ liệu được biến đổi để đáp ứng yêu cầu của hệ thống đích. Điều này có thể bao gồm việc làm sạch dữ liệu, chuyển đổi định dạng, hay thậm chí là tính toán các chỉ số mới.
3. **Load (Lữu trữ):** Dữ liệu đã được biến đổi được nạp vào hệ thống lưu trữ, thường là một kho dữ liệu hoặc data warehouse.

Data Pipe:

Đường ống dữ liệu là một khái niệm quan trọng trong triển khai dòng dữ liệu hiệu quả. Được xây dựng trên ý tưởng của việc tự động hóa quá trình chuyển động và biến đổi dữ liệu, các đường ống dữ liệu đóng vai trò như các luồng làm việc mạnh mẽ.

Thông qua việc sử dụng đường ống dữ liệu, các tỷ lệ lớn dữ liệu có thể được xử lý một cách linh hoạt và hiệu quả. Các tác vụ như xử lý lỗi, theo dõi hiệu suất, và thậm chí là triển khai các biến đổi mới có thể được thực hiện một cách tự động, giúp giảm thiểu sự can thiệp thủ công và tăng tính ổn định của hệ thống.

Thách thức:

- **Tính Nhất Quán của Dữ Liệu:** Đảm bảo tính nhất quán của dữ liệu qua các giai đoạn khác nhau của dòng dữ liệu có thể là thách thức, đặc biệt là trong các hệ thống phân tán.
- **Độ Trễ:** Dòng dữ liệu thời gian thực đòi hỏi độ trễ thấp, điều này có thể là một thách thức trong môi trường cụ thể.
- **Xử lý Lỗi:** Xử lý lỗi trong quá trình dòng dữ liệu là quan trọng để đảm bảo chất lượng và đáng tin cậy của dữ liệu.
- **Khả Năng Mở Rộng:** Khi dung lượng dữ liệu tăng, việc mở rộng các quy trình dòng dữ liệu trở nên quan trọng để duy trì hiệu suất.

Ở mức độ cơ bản, dòng dữ liệu đóng một vai trò quan trọng trong quản lý và xử lý dữ liệu trong các hệ thống khác nhau, và việc hiểu và triển khai nó một cách hiệu quả là quan trọng đối với việc xây dựng các kiến trúc dữ liệu mạnh mẽ hỗ trợ nhu cầu của ứng dụng và doanh nghiệp hiện đại.

2.2.2 Data Orchestration:

Data Orchestration (Điều phối Dữ liệu) là quá trình phối hợp và quản lý nhiều quy trình dữ liệu, quy trình làm việc hoặc dịch vụ khác nhau để đạt được một kết quả cụ thể.

Khái niệm cơ bản:

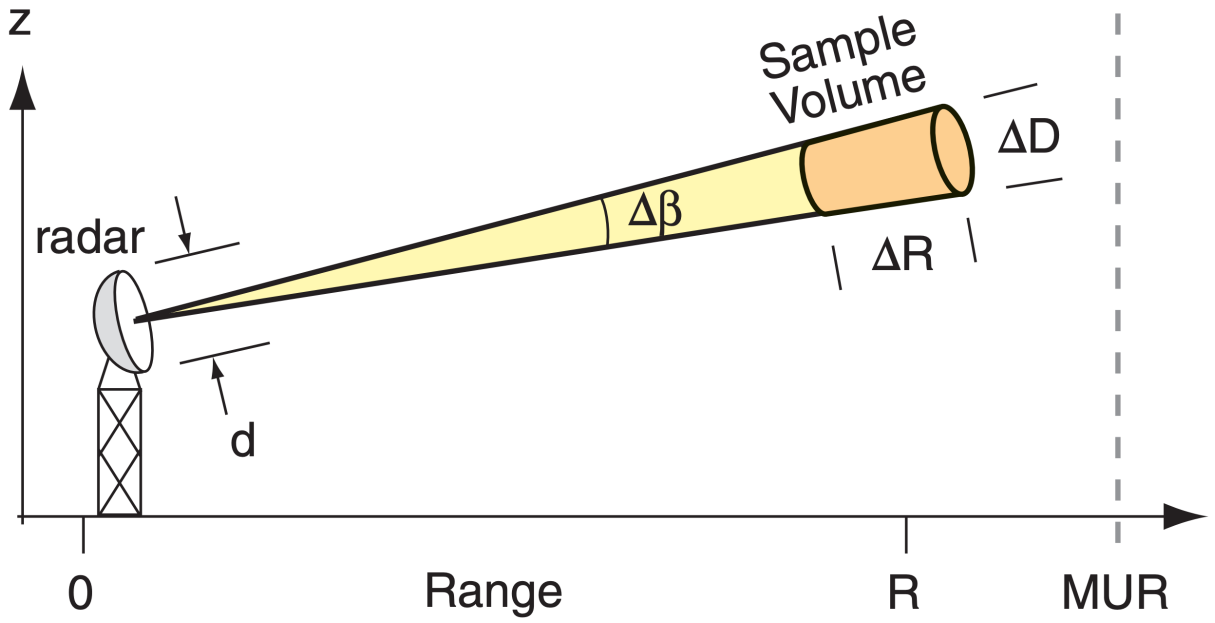
- **Quản lý Quy trình làm việc:** Data Orchestration bao gồm việc định nghĩa và quản lý các quy trình làm việc, xác định thứ tự và sự phụ thuộc giữa các quy trình dữ liệu khác nhau.
- **Lập lịch Công việc:** Các bộ điều khiển lập lịch và thực hiện các công việc vào thời gian phù hợp và theo thứ tự đúng để đạt được kết quả mong muốn.
- **Quản lý Sự phụ thuộc:** Các bộ điều khiển quản lý sự phụ thuộc giữa các công việc, đảm bảo rằng một công việc chỉ được thực hiện khi các công việc phụ thuộc của nó được đáp ứng.
- **Giám sát và Ghi log:** Hệ thống Điều phối cung cấp các công cụ để giám sát tiến trình của các quy trình làm việc và ghi thông tin liên quan để giải quyết sự cố.
- **Phân rã Công việc:** Các bộ điều khiển có thể tối ưu hóa hiệu suất bằng cách phân rã công việc thành nhiều nhiệm vụ và thực hiện chúng song song, cải thiện hiệu quả tổng thể.

Thách thức:

- **Độ phức tạp:** Quản lý các quy trình làm việc phức tạp với nhiều sự phụ thuộc và logic có điều kiện có thể là một thách thức.
- **Môi trường Phân tán:** Việc Điều phối các quy trình dữ liệu trong môi trường phân tán đòi hỏi xử lý các vấn đề như sự cố mạng và sự cố một phần một cách tinh tế.
- **Quản lý Phiên bản:** Quản lý các thay đổi trong quy trình làm việc và đảm bảo tính tương thích ngược khi cập nhật Điều phối có thể là một vấn đề phức tạp.
- **Tải Động:** Điều chỉnh đối với thay đổi động trong khối lượng công việc hoặc nguồn dữ liệu là một thách thức trong Điều phối dữ liệu.

Ở mức độ chi tiết, Data Orchestration đóng một vai trò quan trọng trong việc tối ưu hóa quy trình làm việc dữ liệu và đảm bảo tính hiệu quả của hệ thống. Hiểu và triển khai những nguyên lý này một cách chặt chẽ là quan trọng đối với sự thành công của các ứng dụng và doanh nghiệp trong thời đại số ngày nay.

2.3 Lý thuyết về khí tượng



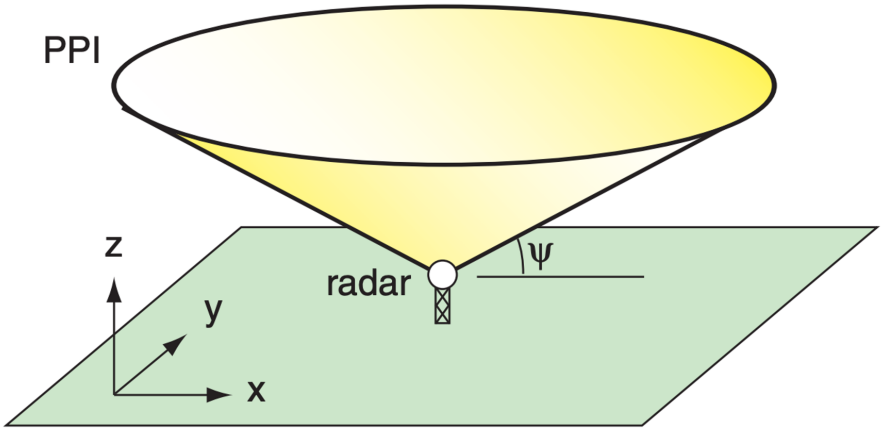
Hình 2.1: *Hệ thống radar thời tiết - [7]*

2.3.1 Các Khái Niệm Cơ Bản

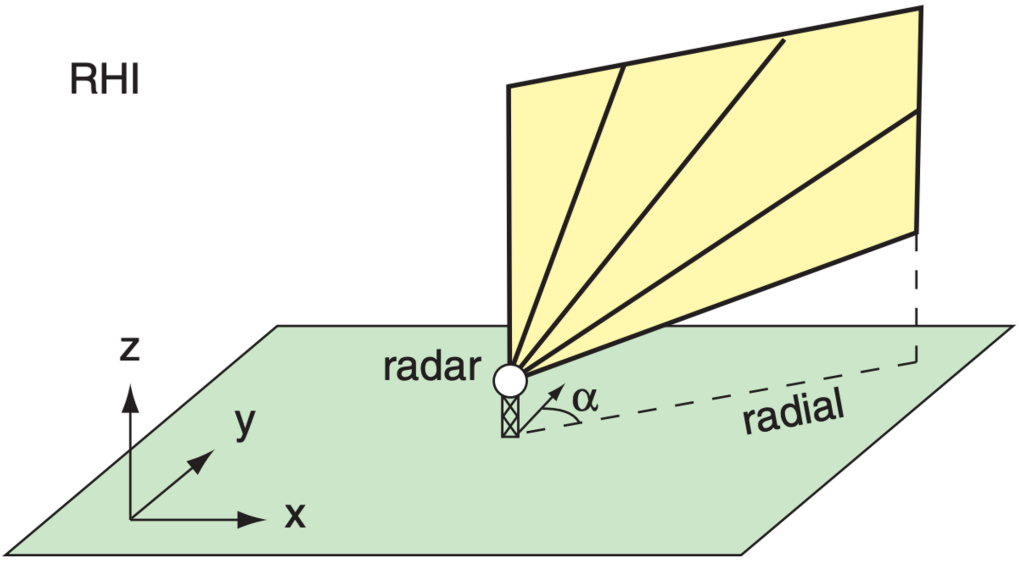
2.3.1.1 Radar thời tiết

[1] Radar thời tiết là một loại cảm biến có khả năng phát sóng vô tuyến (bước sóng trong phạm vi từ 250 - 1000 kW) [7]. Để gia tăng cường độ sóng, một chảo antent (attenna dish) hình parabol được sử dụng nhằm hội tụ bước sóng. Radar có thể nâng và hạ (tuỳ theo yêu cầu) để thu nhập thông tin tại các vị trí chỉ định trong không gian 3 chiều.

Thông thường, các radar được lập trình để quét theo góc hướng (azimuth) 360^o, mỗi vòng sẽ quét ở một gốc nâng khác nhau. Như vậy, radar sẽ mất khoảng từ 4 đến 10 phút để hoàn thành một lần quét.



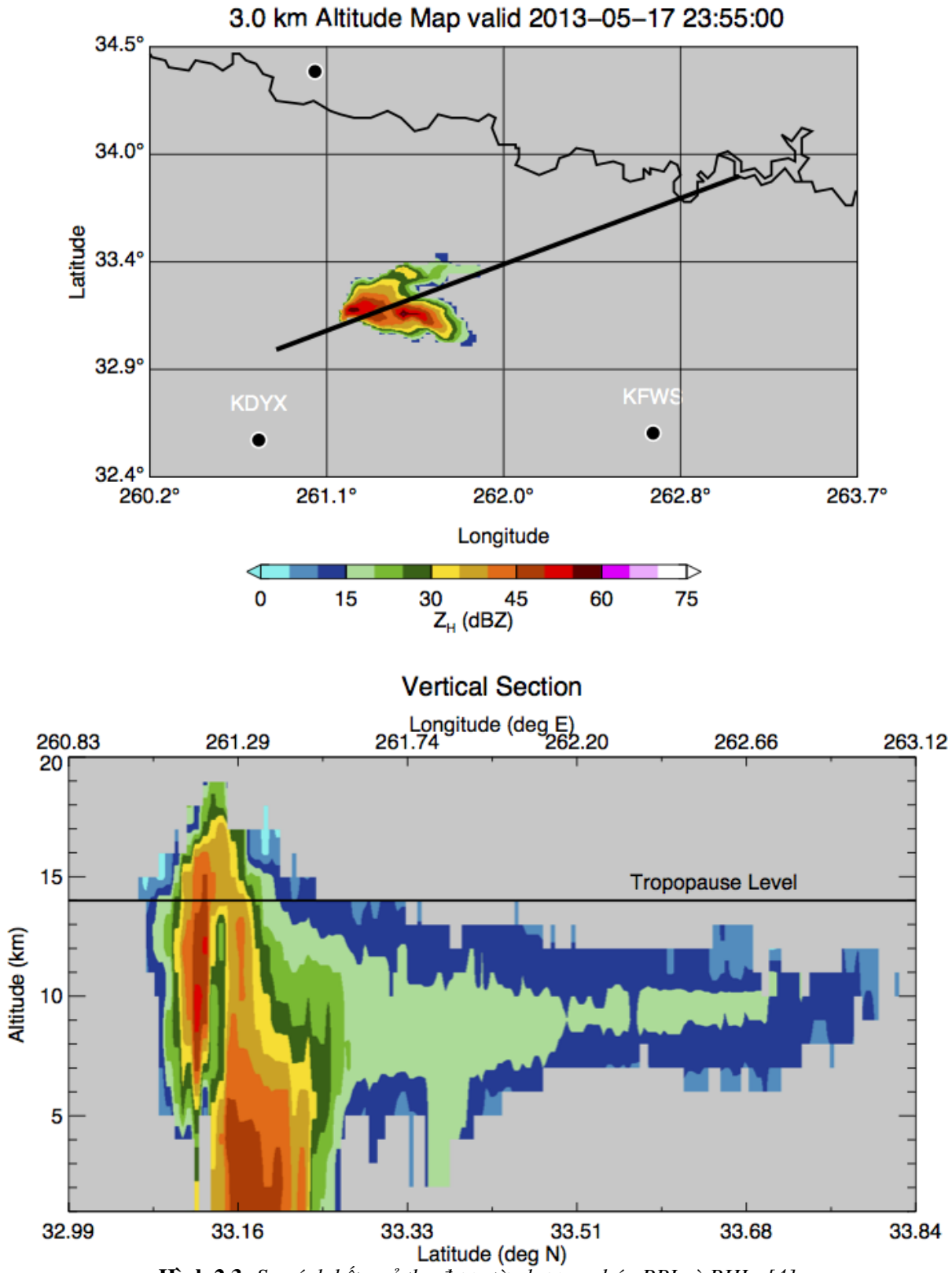
(a) *Sản phẩm quét tròn với góc nâng cố định (Plan-Position Indicator - PPI) - [7]*



(b) *Sản phẩm quét thẳng đứng ở một gốc phương vị nhất định (Range Height Indicator) - [7]*

Đối với biểu diễn PPI, radar sẽ quét toàn bộ góc hướng, nhưng chỉ ở một góc nâng nhất định. Kết quả thu được tương tư một bản đồ trên mặt phẳng. Với RHI, radar sẽ giữ nguyên góc hướng nhưng thay đổi về góc nâng. Kết quả thu được giúp người xem có cái nhìn rõ nét hơn về chiều cao, kích thước của hiện tượng khí tượng.

[1]Tên tiếng Việt của các thuật ngữ sẽ được căn cứ dựa trên TCVN 12636-12 : 2021 [5]



Hình 2.3: *So sánh kết quả thu được từ phương pháp PPI và RHI - [4]*

2.3.1.2 Phương trình Radar và độ phải hồi vô tuyến

Tại một thời điểm, radar sẽ phát ra một luồng sóng trong khoảng thời gian ngắn ($\Delta t = 0.5 - 10\mu s$). Lúc này, tuỳ thuộc mật độ các phân tử tự do trong không khí (hơi nước, khói bụi, ...), năng lượng của bước sóng này sẽ bị hấp thụ một phần. Cường độ bước sóng mà radar nhận được sẽ nhỏ hơn cường độ sóng ban đầu. Tỉ lệ này được thể hiện thông qua **Phương trình radar** [7]:

$$\left[\frac{P_R}{P_T}\right] = [b] \cdot \left[\frac{|K|}{L_a}\right]^2 \cdot \left[\frac{R_1}{R}\right]^2 \cdot \left[\frac{Z}{Z_1}\right]$$

Trong đó, các biến của phương trình gồm có:

- $|K|$ không có đơn vị:
  - $|K|^2 \approx 0.93$ cho các hạt nước lỏng
  - $|K|^2 \approx 0.208$ cho tinh thể băng
- $R(\text{km})$: khoảng cách
- $R_1 = \sqrt{Z_1 \cdot c \cdot \Delta t / \lambda^2}$: hệ số khoảng cách
- Z: Hệ số phản hồi vô tuyến của Radar
- $Z_1 = 1\ \text{mm}^6\ \text{m}^{-3}$: hệ số đơn vị phản hồi vô tuyến

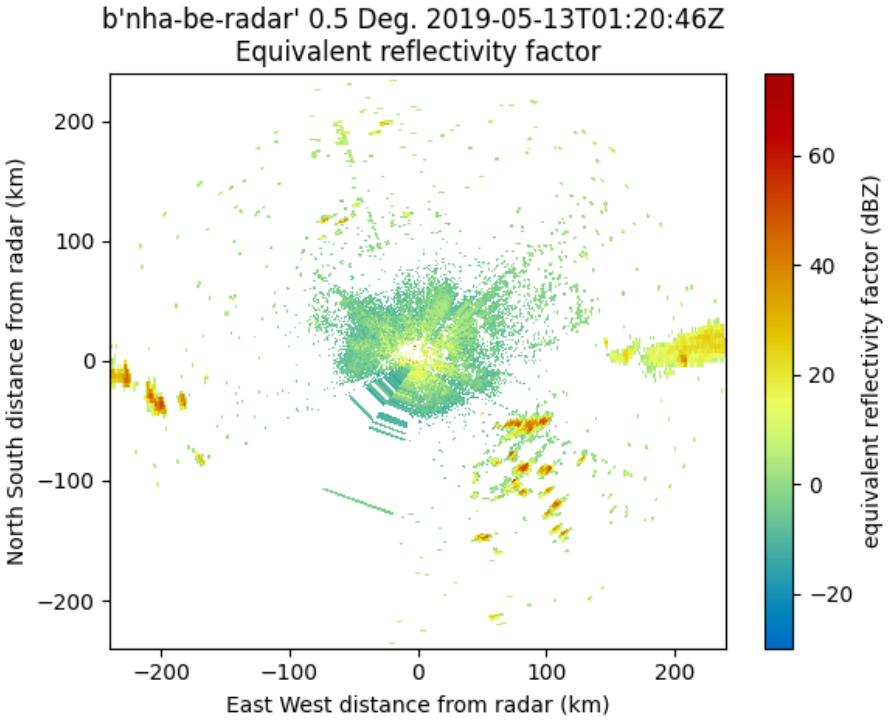
Từ phương trình Radar, ta suy ra được công thức tính độ phản hồi vô tuyến:

$$\text{dBZ} = 10\left[\log\left(\frac{P_R}{P_T}\right) + 2\log\left(\frac{R}{R_1}\right) - 2\log\left|\frac{K}{L_a}\right| - \log(b)\right]$$

Các nhà khí tượng thuỷ văn học thường quan tâm đến con số này vì nó tỉ lệ thuận với mức độ giáng thuỷ (precipitation).

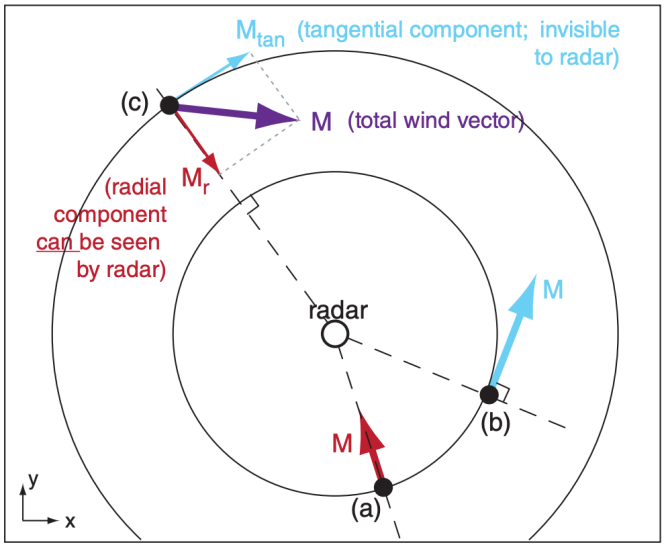
Giá trị (dBZ)	Thời tiết
-28	Sương mù
-12	Không khí trong lành
25 - 30	Tuyết khô / mưa nhẹ
40 - 50	Mưa lớn
75	Mưa đá khổng lồ

Bảng 2.1: *Tương quan hệ số phản xạ của radar và giáng thuỷ - Stull [7]*



Hình 2.4: *Minh hoa hệ số phản xạ từ dữ liệu radar Nhà Bè*

2.3.1.3 Vận tốc xuyên tâm



Hình 2.5: *Minh hoạ các tình huống vận tốc mà radar Doppler có thể quan sát. (a) Phương của gió trùng tại M trùng với đường kính đường tròn có tâm tại radar, radar có thể xác định được vận tốc tại đây. (b) Phương của gió trùng với tiếp tuyến của đường tròn, radar không thể xác định được vận tốc. (c) Phân tích hướng gió tại M thành 2 vận tốc vuông góc nhau, radar chỉ xác định được vector vận tốc theo M_r*

Khi các sóng vô tuyến từ các radar Doppler này truyền đến các phân tử trong không khí, sự chuyển dịch vị trí của các hạt này làm lệch pha giữa tín hiệu truyền đi và tín hiệu nhận lại được. Các radar sẽ căn cứ vào thông tin này để tính toán vận tốc gió tại các điểm trong không gian.

2.4 Định dạng dữ liệu trong phân tích khí tượng

2.4.1 Định dạng SIGMET - raw format (Vaisala)

Vaisala là một công ty Phần Lan, chuyên về các lĩnh vực thuộc mội trường và khí tượng thuỷ văn. Định dạng RAW (trong một số tài liệu còn gọi là SIGMET [1]) là một trong những định dạng lưu trữ mà công ty này đã phát triển ra nhằm thực hiện tổ chức dữ liệu xuất ra từ các thiết bị radar của họ.

Một số điểm nổi bật về định dạng này có thể kể đến như:

- Nội dung của file được phân thành một **block** lớn. Mỗi block có kích thước đúng 6144 bytes. Kích thước này vừa đúng bằng kích thước lưu trữ chính trên các thiết bị băng từ cũ.
- File thường là tổng hợp của tất cả những lần radar tiến hành quét dữ liệu.
- Các phần dữ liệu (record) sẽ được sắp xếp trong phạm vi 1 block (6144 bytes). Trong trường hợp phần block còn dư, dữ liệu sẽ được đệm thêm các số 0.

Với những đặc điểm kể trên, có thể nhận thấy các ưu điểm chính từ việc lưu trữ định dạng RAW bao gồm: [8]

- Thân thiện với các loại băng từ. Đây thường là các thiết bị phổ biến trước đây, hiện nay vẫn được sử dụng rộng rãi nhờ vào hiệu quả từ mức dung lượng / chi phí.
- Nhờ sử dụng cơ chế block, SIGMET giúp các hệ thống lưu trữ thực hiện các biện pháp hồi phục (error recovery) trên mức block.

Nhược điểm chính mà nhóm quan ngại là khả năng mapping (liên kết) giữa cấu trúc khi lưu trữ trong ổ cứng và trên băng từ.

2.4.2 Định dạng NETCDF - Network Common Data Form

NetCDF (Network Common Data Form) là một định dạng tệp tin linh hoạt được thiết kế chặt chẽ để lưu trữ dữ liệu khoa học đa chiều. Trong hệ thống thư viện netCDF, có nhiều định dạng nhị phân được hỗ trợ, mỗi định dạng đóng góp vào tính linh hoạt và khả năng mở rộng của quản lý dữ liệu [6]. Đáng chú ý, các định dạng này bao gồm:

1. Định dạng Classic: Ban đầu được sử dụng trong phiên bản đầu tiên của netCDF và vẫn là lựa chọn mặc định cho việc tạo tệp tin.
2. Định dạng 64-bit Offset: Giới thiệu từ phiên bản 3.6.0, định dạng này hỗ trợ kích thước biến và tệp tin lớn hơn.
3. Định dạng netCDF-4/HDF5: Xuất hiện từ phiên bản 4.0, sử dụng định dạng dữ liệu HDF5 với một số hạn chế.
4. Định dạng HDF4 SD: Hỗ trợ chủ yếu cho việc đọc dữ liệu.
5. Định dạng CDF5: Hỗ trợ được đồng bộ với dự án parallel-netcdf.

Tất cả các định dạng này đều thể hiện tính tự mô tả, với một phần tiêu đề chi tiết mô tả cấu trúc của tệp tin, bao gồm các mảng dữ liệu và siêu dữ liệu tệp tin dưới dạng thuộc tính tên/giá trị. Thiết kế này đảm bảo tính độc lập với nền tảng, với các vấn đề như endianness được giải quyết một cách linh hoạt thông qua các thư viện phần mềm.

Hãy xem xét ví dụ cụ thể về việc lưu trữ các thông số khí tượng quan trọng như nhiệt độ, độ ẩm, áp suất, tốc độ và hướng gió trong các tệp tin netCDF. Điều này minh họa khả năng của định dạng này trong xử lý các bộ dữ liệu khoa học đa dạng, cung cấp một phương tiện mạnh mẽ và linh hoạt để quản lý thông tin đa chiều.

```
→  titan2023 ncdump -h radar.nc
netcdf radar {
dimensions:
        time = UNLIMITED ; // (1748 currently)
        range = 1198 ;
        sweep = 5 ;
        string_length = 32 ;
```

Hình 2.6: *Thông tin radar ở định dạng NETCDF. Số chiều của bộ dữ liệu tổng cộng là 2975 chiều, được phân nhóm cho 4 nhãn khác nhau.*

Bắt đầu từ phiên bản 4.0, API netCDF giới thiệu khả năng sử dụng định dạng dữ liệu HDF5. Sự tích hợp quan trọng này cho phép người dùng netCDF tạo tệp tin HDF5, mở khóa những lợi ích như kích thước tệp tin lớn hơn đáng kể và hỗ trợ cho nhiều chiều không giới hạn. Bước tiến này đánh dấu một bước quan trọng hướng tới việc tận dụng những ưu điểm mở rộng của định dạng HDF5.

NetCDF Classic và Định dạng 64-bit Offset là tiêu chuẩn quốc tế của Open Geospatial Consortium[2], thể hiện sự chắc chắn và độ tin cậy trong việc đảm bảo khả năng tương thích và mở rộng của định dạng netCDF trên toàn cầu.

2.5 Công nghệ sử dụng

Chương 3

Phân tích và thiết kế hệ thống

3.1 Tổng quan

3.2 Thăm dò dữ liệu

3.3 Mô hình cơ sở dữ liệu

3.4 Kiến trúc hệ thống

Chương 4

Hiện thực

4.1 Luồng dữ liệu

Phần hiện thực của nhóm sẽ nằm trong năm bước còn lại trong mô tả tại hình **??**.

Tại bước 3, nhóm sẽ setup (cài đặt) một server SFTP đơn giản. SFTP là một giao thức đơn giản và phổ biến. Hiện nay, có rất nhiều những thư viện và công cụ để hỗ trợ giao tiếp dựa trên giao thức này. Ngoài ra, so với FTP, giao thức kể trên còn đảm bảo tính bảo mật trong suốt quá trình chuyển dịch dữ liệu. Tuỳ thuộc vào mức độ cho phép, nhóm có thể hỗ trợ phía trạm quan trắc xây dựng các scripts (kịch bản) để tự động forward (chuyển tiếp) các file sau khi đã được xử lý tại đây. Hoặc ngược lại, phía trạm quan sát có thể gửi file đến server trên một cách thủ công.

Khi file đã được upload đến server SFTP, nhóm sử dụng Airflow để điều hành tất cả các luồng ETL hiện có trong hệ thống chung. Ở thời điểm hiện tại, nhóm chỉ dừng lại với một DAG duy nhất, để xử lí dữ liệu đến từ trạm quan trắc Nhà Bè. Airflow sẽ tiến hành quan sát những file được thêm mới vào server SFTP của chúng ta và khởi chạy ETL. Việc lựa chọn Apache Spark ở đây dựa trên khối lượng dữ liệu và độ phức tạp được đặt ra. Nếu lượng dữ liệu là không quá nhiều cho mỗi file SIGMET mới, và bản thân Python có xử lí được, không cần thiết phải sử dụng Spark ở bước này.

Dữ liệu về khí tượng khi được đưa đến cơ sở hạ tầng của nhóm sẽ được phân ra hai luồng chính: Những metadata (thông tin mở rộng) của dữ liệu gốc như ngày tạo ra, kích thước, thời điểm ghi nhận, ... sẽ được lưu trong một RDBMS (hệ quản trị cơ sở dữ liệu quan hệ) truyền thống. Cụ thể ở đây, nhóm lựa chọn PostgreSQL nhờ vào độ phổ biến và mức độ am hiểu của nhóm. Các metadata lưu trữ ở đây giúp cơ sở dữ liệu của nhóm nhanh chóng phản hồi các query (truy vấn) mà chưa cần trực tiếp phải sử dụng đến dữ liệu gốc. Một số query phổ biến có thể kể đến như:

- Các mốc thời gian đang được ghi nhận bao gồm những gì? (Ví dụ: từ ngày 21/11/2023 cho đến ngày 17/12/2023)
- Tại thời điểm x, radar có toạ độ địa lý là bao nhiêu?
- Các trường dữ liệu đang được lưu trữ là gì?

Bên cạnh đó, DB (cơ sở dữ liệu) trên còn đóng vai trò như mục index (chỉ mục) giúp hệ thống nhanh chóng xác định vị trí lưu trữ của dữ liệu gốc.

Với các dữ liệu về khí tượng thuỷ văn cụ thể, nhóm nhận thấy rằng sẽ không thật sự hiệu quả khi lưu trữ chúng trực tiếp trên các DBMS trên. Đồng thời, nhóm nhận thấy việc lưu trữ dữ liệu trên file vẫn đem đến một kích thước tổng quan hợp lý. Vì vậy, nhóm quyết định sẽ tách phần dữ liệu thô ra và lưu trữ trực tiếp trên các files. Đồng thời kết hợp với các index (đã đề cập ở trên) để tăng tốc quá trình truy xuất.

Để tạo cửa ngõ cho việc truy vấn dữ liệu, phục vụ cho các bên về model, machine learning và AI, ... nhóm sẽ phát triển một server Backend đơn giản, sử dụng FastAPI của Python để giúp tăng tốc độ phát triển giải pháp. Tại bước 5, backend sẽ nhận dữ liệu truy vấn dưới định dạng REST API (tại bước 6), truy vấn dữ liệu trong DB của metadata và trong các file dữ liệu và và trả về kết quả đạt được. Ở những lần train khác nhau, các bên của Machine Learning có thể kết nối đến server này để lấy dữ liệu.

Cần nói thêm, toàn bộ hệ thống sẽ được phát triển và vận hành theo hướng containerize (đóng gói) và sẽ được deploy (triển khai) trên nền tảng Kubernetes. Việc này thể hiện khả năng của hệ thống trong việc duy trì tính sẵn sàng cao (High-Availability) cũng như dễ dàng trong việc duy trì giải pháp. Trong phạm vi phần minh hoạ này, nhóm sẽ chỉ dừng lại với việc triển khai trên một cụm máy tính nhúng Raspberry Pi.

Sau cùng, tại bước 7, nhóm muốn đề xuất thêm một vấn đề. Nếu phù hợp, nhóm có thể xây dựng thêm một `DataLoader` (bộ nạp dữ liệu) đề phục vụ nhanh chóng đến các nhóm làm model khác. Một trong những thư viện phổ biến hiện nay của các bên AI là Pytorch, nên nhóm sẽ tiếp cận với nền tảng này trước.

Chương 5

Kiểm thử

5.1 Unit testing

5.2 Integrated testing

Trong nghiên cứu này, chúng tôi đã thành công trong việc xây dựng một Proof-of-Concept (chứng minh khái niệm) mang tính ứng dụng cao, nhằm mục đích giảm thiểu các công đoạn trong quy trình làm việc thông thường. Đây là một bước quan trọng để tối ưu hóa và cải thiện hiệu suất làm việc trong các ngữ cảnh nghiên cứu và thực tế.

Chúng tôi đã đặt ra mục tiêu tạo ra một hệ thống linh hoạt có khả năng thích ứng cao, giúp giảm bớt những bước phức tạp trong quy trình công việc. Bằng cách này, chúng tôi không chỉ giúp tăng cường hiệu suất mà còn giảm áp lực công việc đối với nhân sự, tạo điều kiện thuận lợi cho sự sáng tạo và tập trung vào các nhiệm vụ chính.

Chúng tôi không chỉ dừng lại ở việc phát triển hệ thống mà còn đề xuất các chiến lược triển khai linh hoạt, nhấn mạnh sự tích hợp dễ dàng vào môi trường làm việc hiện tại của những người đang thực hiện công việc thu thập thông tin và dự báo thời tiết.

Chương 6

Hướng phát triển

Phát triển thành Hệ thống nền tảng dữ liệu thời tiết

Hệ thống nền tảng dữ liệu thời tiết (Weather Data Platform - WDP) được phát triển với mục tiêu trở thành một giải pháp toàn diện cho việc khai thác sức mạnh của dữ liệu thời tiết. Hệ thống được thiết kế để đáp ứng những yêu cầu cụ thể của các nhà khí tượng, học giả, nghiên cứu học thuật và các nhà phát triển từ nhiều lĩnh vực khác nhau, bao gồm cả freelancers, doanh nghiệp và tổ chức phi chính phủ (Non-governmental Organizations - NGOs). WDP đóng vai trò như một trung tâm tập trung cho việc tích hợp dữ liệu thời tiết, phân tích, và nhiều tính năng khác.

Với mong muốn phát triển thành Hệ thống nền tảng dữ liệu thời tiết, chúng tôi hướng đến sự hoàn thiện và đa chiều hoá thông tin thời tiết. Không chỉ là một bảng số liệu, mà là một trải nghiệm toàn diện. Trong tương lai, bên cạnh việc tiếp tục xây dựng cơ sở dữ liệu tích hợp theo hướng đã đề xuất, chúng tôi hứa hẹn sẽ tiếp tục nghiên cứu để mở rộng và phát triển cơ sở dữ liệu tích hợp này thành hệ thống nền tảng dữ liệu thời tiết với những hướng phát triển như sau:

1. **Dữ liệu phi tuyến:** Mở rộng từ việc tích hợp dữ liệu cơ bản, chúng tôi sẽ chú trọng vào việc cung cấp dữ liệu phi tuyến, chi tiết và đa nguồn, giúp người dùng khám phá thêm về môi trường xung quanh.

2. **Trí tuệ nhân tạo thấu hiểu**: Sử dụng trí tuệ nhân tạo để thấu hiểu ngôn ngữ của thời tiết, từ những biến đổi nhỏ đến những sự kiện lớn, tạo nên một nguồn thông tin thời tiết sâu sắc và thông minh.

3. **Giao diện người dùng tương tác**: Không chỉ là việc truy cập thông tin, mà còn là việc tương tác với dự báo thời tiết. Giao diện người dùng sẽ là nơi người dùng thể hiện sự tò mò và tương tác trực tiếp với dữ liệu.

4. **Kết Nối Thông Tin Địa Lý**: Tận dụng hệ thống thông tin địa lý để mang đến cái nhìn thực tế hóa, địa bàn hóa cho dự báo thời tiết. Điều này giúp người dùng hiểu rõ hơn về tác động thời tiết đối với môi trường xung quanh họ.

5. **Tối ưu hoá hiệu suất**: đảm bảo khả năng đáp ứng nhanh chóng và đồng đều trong mọi điều kiện.

6. **Bảo mật dữ liệu**: Tăng cường an toàn dữ liệu để đảm bảo tính bảo mật và toàn vẹn của thông tin thời tiết.

7. **Hệ thống dự báo nâng cao**: Nghiên cứu và tích hợp trí tuệ nhân tạo để cải thiện khả năng dự báo và đưa ra thông tin dự báo cáo chính xác.
8. **Kiểm thử và tối ưu hoá**: Tiến hành kiểm thử hệ thống để đảm bảo tính ổn định và xử lý mọi vấn đề tiềm ẩn. Tối ưu hóa hiệu suất nếu cần.
9. **Triển khai và duy trì**: Triển khai hệ thống và duy trì một chu kỳ cập nhật đều đặn để đảm bảo rằng nó luôn cung cấp thông tin thời tiết chính xác và đáng tin cậy.

Tài liệu tham khảo

[1] Radxconvert - lrose wiki. URL http://wiki.lrose.net/index.php/RadxConvert.

[2] Ogc standard netcdf classic and 64-bit offset. https://www.opengeospatial.org/standards/netcdf, Accessed: 2017-12-05. Archived from the original on 2017-11-30. Retrieved 2017-12-05.

[3] Tyler Akidau, Robert Bradshaw, Craig Chambers, Slava Chernyak, Rafael J. Fernández-Moctezuma, Reuven Lax, Sam McVeety, Daniel Mills, Frances Perry, Eric Schmidt, and Sam Whittle. The dataflow model: A practical approach to balancing correctness, latency, and cost in massive-scale, unbounded, out-of-order data processing. *Proceedings of the VLDB Endowment*, 8:1792–1803, 2015.

[4] casey. Using range height indicator scan of radar. URL https://earthscience.stackexchange.com/questions/7222/using-range-height-indicator-scan-of-radar. Truy cập lần cuối ngày 17/12/2023.

[5] Tổng cục Khí tượng Thủy văn. Quan trắc khí tượng thuỷ văn - phần 12: Quan trắc ra đa thời tiết. *TCVN 12636-12 : 2021*, 2021.

[6] Russ Rew, Glenn Davis, Steve Emmerson, Cathy Cormack, John Caron, Robert Pincus, Ed Hartnett, Dennis Heimbigner, Lynton Appel, and Ward Fisher. Unidata netcdf, 1989. URL http://www.unidata.ucar.edu/software/netcdf/.

[7] Roland Stull. Weather Radars, 12 2022. [Truy cập lần cuối ngày 17/12/2023].

[8] *RAW Product Format - IRIS Programming Guide - IRIS Radar*. Vaisala. URL https://ftp.sigmet.vaisala.com/files/html_docs/IRIS-Programming-Guide-Webhelp/raw_product_format.html.