VIETNAM NATIONAL UNIVERSITY, HO CHI MINH CITY

HO CHI MINH UNIVERSITY OF TECHNOLOGY

FACULTY OF COMPUTER SCIENCE AND ENGINEERING

CAPSTONE PROJECT REPORT

BUILDING AN INTEGRATED DATABASE FOR SHORT TERM FORECASTING SYSTEMS IN HYDRO-METEOROLOGY

Major: Computer Science

THESIS COMMITTEE:	**Council 9**
SUPERVISORS:	**Lê Hồng Trang**
	Trương Quỳnh Chi
REVIEWER:	**Trương Quỳnh Chi**
	—o0o—
STUDENT 1:	**Trần Hà Tuấn Kiệt (2011493)**
STUDENT 2:	**Nguyễn Đức Thuỵ (2012158)**

Ho Chi Minh City, 6/2024

Declaration

We guarantee that this research is our own, conducted under the supervision and guidance of Assoc. Prof. Dr. Lê Hồng Trang. The result of our research is legitimate and has not been published in any form prior to this. All materials used within this research are collected by ourselves, by various sources and are appropriately listed in the references section. In addition, within this research, we also used the results of several other authors and organizations. They have all been aptly referenced. In any case of plagiarism, we stand by our actions and are to be responsible for it. Ho Chi Minh City University of Technologytherefore is not responsible for any copyright infringements conducted within our research.

Ho Chi Minh City, June 3, 2024

Team Members

Trần Hà Tuấn Kiệt

Nguyễn Đức Thuỵ

Acknowledgments

We, the research group consisting of two members, would like to send our thanks toward everyone who has contributed to the success of this thesis.

First and foremost, we would like to express our gratitude to Assoc. Prof. Dr. Lê Hồng Trang . The dedication and extensive knowledge of our supervisor not only guided us through the challenges of the project but also helped us develop research skills and critical thinking.

We also want to extend special thanks to all the friends, colleagues, and family who supported us throughout the research process. The input and opinions of everyone enriched the content and quality of the project.

Last, but certainly not least, we appreciate each other - research partners accompanying us in every step of the project. Our collaboration and joint contributions have resulted in a product that we take pride in.

We believe that this project marks a significant step in our development and could not have been achieved without the support and contributions of everyone involved.

Abstract

In this document, we introduce a comprehensive proposal to integrate, coordinate, and monitor meteorological and hydrological data in Vietnam. The proposed system can serve as a foundation for building a national database specializing in meteorological information.

To clarify our proposal, we have built a comprehensive pilot version, in which we have simulated the process of collecting and transmitting data from the weather station at Nha Be. At the same time, we focused on the process of converting and organizing data storage to ensure transparency and optimal performance in information processing.

This research is a manifestation of significant efforts in optimizing the management of meteorological data in Vietnam. The proposed system not only addresses the challenges of integration and coordination but also lays the foundation for a robust national database, serving the diverse needs of research and applications in the field of meteorology.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem statement

In recognizing the need for National Meteorological Services (NMHSs) to improve their climate data and monitoring services, there is a deliberate focus on those aspects of climate data management wishing to make the transition to a modern climate database management system and, just as important, on what skills, systems and processes need to be in place to ensure that operations are sustained. In the context of the ever-growing complexity of climate change, the task of creating an integrated database for short-term forecasting systems in the fields of meteorology and hydrology poses a considerable challenge. The question at hand is how we can optimize the management of weather information from multiple sources and store it efficiently in a database. This optimization is crucial to ensure the provision of synchronized and high-quality information to support forecasting systems.

1.2 Motivation

Information about the weather has been recorded in manuscript form for many centuries. The early records included notes on extreme and, sometimes, catastrophic events and also on phenomena such as the freezing and thawing dates of rivers, lakes and seas, which have taken on a higher profile with recent concerns about climate change.

Specific journals for the collection and retention of climatological information have been used over the last two or three centuries (WMO 2005). The development of instrumentation to quantify meteorological phenomena and the dedication of observers to maintaining methodical,

reliable and well-documented records paved the way for the organized management of climate data. Since the 1940s, standardized forms and procedures gradually became more prevalent and, once computer systems were being used by NMHSs, these forms greatly assisted the computerized data entry process and consequently the development of computer data archives. The latter part of the twentieth century saw the routine exchange of weather data in digital form and many meteorological and related data centers took the opportunity to directly capture and store these in their databases. Much was learned about automatic methods of collecting and processing meteorological data in the late 1950s, a period that included the International Geophysical Year and the establishment of the World Weather Watch. The WMO's development of international guidelines and standards for climate data management and data exchange assisted NMHSs in organizing their data management activities and, less directly, also furthered the development of regional and global databases. Today, the management of climate records requires a systematic approach that encompasses paper records, microfilm/microfiche records and digital records, where the latter include image files as well as the traditional alphanumeric representation.

Before electronic computers, mechanical devices played an important part in the development of data management. Calculations were made using comptometers, for example, with the results being recorded on paper. A major advance occurred with the introduction of the Hollerith system of punch cards, sorters and tabulators. These cards, with a series of punched holes recording the values of the meteorological variables, were passed through the sorting and tabulating machines enabling more efficient calculation of statistics. The 1960s and 1970s saw several NMHSs implementing electronic computers and gradually the information from many millions of punched cards was transferred to magnetic tape. These computers were replaced with increasingly powerful mainframe systems and data were made available online through developments in disk technology.

Aside from advances in database technologies, more efficient data capture was made possible through the mid-to-late 1990s with an increase in automatic weather stations (AWSs), electronic field books (i.e. on-station notebook computers used to enter, quality control and transmit observations), the Internet and other advances in technology. Not surprisingly, there are a number of trends already underway that suggest there are many further benefits for NMHSs in managing data and servicing their clients. The Internet is already delivering greatly improved data access capabilities and, providing security issues are managed, we can expect major op-

portunities for data managers in the next five to ten years. In addition, Open Source7 relational database systems may also remove the cost barriers to relational databases for many NMHSs over this period.

1.3 Project Goal

It is essential that both the development of climate databases and the implementation of data management practices take into account the needs of the existing, and to the extent that it is predictable, future data users. While at first sight this may seem intuitive, it is not difficult to envisage situations where, for example, data structures have been developed that omit data important for a useful application or where a data centre commits too little of its resources to checking the quality of data for which users demand high quality.

In all new developments, data managers should either attempt to have at least one key data user as part of their project team or undertake some regular consultative process with a group of user stakeholders. Data providers or data users within the organization may also have consultative processes with end users of climate data (or information) and data managers should endeavour to keep abreast of both changes in needs and any issues that user communities have. Put simply, data management requires awareness of the needs of the end users.

At present, the key demand factors for data managers are coming from climate prediction, climate change, agriculture and other primary industries, health, disaster/emergency management, energy, natural resource management (including water), sustainability, urban planning and design, finance and insurance. Data managers must remain cognizant that the existence of the data management operation is contingent on the centre delivering social, economic and environmental benefit to the user communities it serves. It is important, therefore, for the data manager to encourage and, to the extent possible, collaborate in projects which demonstrate the value of its data resource. Even an awareness of studies that show, for example, the economic benefits from climate predictions or the social benefits from having climate data used in a health warning system, can be useful in reminding senior NMHS managers or convincing funding agencies that data are worth investing in. Increasingly, value is being delivered through integrating data with application models (e.g. crop simulation models, economic models) and so integration issues should be considered in the design of new data structures.

1.4 Project Scope

This project focuses on Ho Chi Minh City, a sprawling metropolis with a distinct climate that significantly impacts the daily lives of its residents. To provide the most valuable weather insights, we will implement a two-pronged approach.

Firstly, We will conduct research and collect data from the meteorological and hydrological monitoring stations in the city to ensure diversity and representation of local weather conditions. At the same time, we will proceed with preprocessing and standardizing the data to ensure accuracy and consistency of the dataset. Based on what we have collected, we will populate the data to proper formats and structures.

Secondly, this project entails the development of a comprehensive data platform. This platform will integrate a central database, designed to efficiently store and manage weather information from diverse sources. By consolidating these disparate data streams, we aim to create a unified and dependable resource for weather data. This platform is anticipated to significantly enhance the workflow of end-users and will be built with scalability in mind to accommodate future growth in data volume and user demand.

1.5 Stakeholders

1.5.1 The National Center for Hydro-meteorological Forecasting

The National Center for Hydrometeorological Forecasting (NCHMF), abbreviated for "Trung tâm Dự báo khí tượng thuỷ văn quốc gia" in Vietnamese, is an organizational unit under the General Department of Meteorology and Hydrology, Ministry of Natural Resources and Environment[4]. The National Hydro-Meteorological Forecasting Center has several crucial missions, including the establishment and presentation of standards and technical regulations for meteorological and hydrological forecasting, the operation of the national forecasting and warning system, monitoring and reporting on weather conditions and climate change, issuing and disseminating forecast bulletins and warnings, and participating in international meteorological agreements. Additionally, the center is responsible for conducting research, application, and technology transfer related to forecasting and warning, and implementing administrative reform and anti-corruption measures. These key missions contribute significantly to the center's

role in ensuring public safety and providing timely essential meteorological and hydrological information.

1.5.2 Nha Be Weather Radar Station

The Nha Be weather radar station is located on Nguyen Van Tao Street, Long Thoi Commune, Nha Be District, Ho Chi Minh City. This radar station has been in operation since 2004 and is managed by the Southern Region Hydro-Meteorological Center. Its mission is to monitor and supervise weather phenomena within a radius of 480 km from the center of Ho Chi Minh City, and to provide warnings and forecasts for dangerous weather conditions such as storms, tropical depressions, and thunderstorms within a radius of 300 km. Additionally, the station serves the purpose of forecasting rain and heavy rain for the Ho Chi Minh City area within a radius of approximately 120 km.

The Nha Be weather radar station is of the Doppler DWSR-2500C type, operating on the C-band frequency. It is manufactured by the U.S. company EEC and has the capability to scan signals within a radius of 300 km. In addition to being the forecasting center for the entire Southern region and Ho Chi Minh City, the Nhà Bè radar station also has the task of predicting rain and providing data for the flood control center of Ho Chi Minh City.

1.6 Solutions

MeteorFlow is a cloud-based platform for managing and analyzing weather data. Key features include real-time data access from a global network, advanced tools for weather pattern analysis, automation capabilities for data collection and processing, scalable infrastructure for handling large data volumes, secure storage, and community support from meteorological experts. These functionalities streamline workflows, allowing for faster, data-driven decisions in meteorology.

For more detailed insights and to explore their solutions, please visit the website:

`https://www.demo.meteorflow.com/`

Chapter 2

Theoretical basis

2.1 Data system

2.1.1 Traditional Database Systems and Database Management Systems

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, the 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 has 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 the database approach are identity, attributes and relationships. An entity is something that can be identified in the user's 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 the 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 is the main concept 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, independent of the programs that access it [10]. 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 [3].

The 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 the 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 the 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 program and data independence. DBMS access programs do not require changes when the structure of the data files is 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 perspectives of databases depending 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 the

database for modification. Concurrency control is crucial for a DBMS so that the results of the updates are correct. The DBMS software ensures that concurrent transactions operate correctly when several users are trying to update the same data.

Using a DBMS also eliminates unnecessary data redundancy. In the database approach, each primary fact is generally recorded in only one place in the database. 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 to prohibit inconsistencies among files. By eliminating data redundancy inconsistencies among data are also reduced [3]. Reducing redundancy improves the consistency of data while reducing the waste in storage space. DBMS allows 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 to regulate which parts of the database various users can access or update.

2.1.2 A Modern Approach to Data Systems

In contemporary computing environments, the dominance has shifted towards data-intensive applications, deviating from the traditional emphasis on compute-intensive tasks. The limiting factor for these applications seldom resides in the sheer computational power of the CPU; rather, the primary challenges typically revolve around the magnitude of the data, its intricate structures, and the rapidity with which it changes. Unlike compute-intensive operations that heavily rely on processing speed, data-intensive applications, dealing with extensive datasets, intricate data structures, or swiftly evolving information, necessitate adept strategies for storage, retrieval, and manipulation. Consequently, effectively addressing the multifaceted dynamics of data becomes paramount, highlighting the imperative for sophisticated data management and processing techniques to optimize performance in the face of these intricate challenges.

Why should we amalgamate these diverse elements within the overarching label of data systems? Recent years have witnessed the emergence of a plethora of novel tools for data storage and processing, each meticulously optimized for an array of distinct use cases [16]. Consider,

for instance, datastores that concurrently function as message queues (e.g., Redis) or message queues equipped with database-like durability assurances (such as Apache Kafka). The demarcation lines between these categories are progressively fading, reflecting a landscape where boundaries are increasingly ambiguous.

Moreover, a growing number of applications now present challenges of such magnitude or diversity that a solitary tool is no longer sufficient to fulfill all its data processing and storage requisites. Instead, the workload is deconstructed into tasks amenable to efficient execution by individual tools. These disparate tools are then intricately interwoven using application code, offering a nuanced and adaptable approach to the multifaceted demands of contemporary data management and processing.

In navigating the complex landscape of application development, the quest for reliability, scalability, and maintainability unveils a challenging yet essential journey. As we delve into the intricate patterns and techniques that permeate various applications, we embark on an exploration to fortify these foundational pillars in the realm of software systems.

Ensuring reliability involves the meticulous task of ensuring systems function correctly, even when confronted with faults. These faults may manifest in the hardware domain as random and uncorrelated issues, in software as systematic bugs that are challenging to address, and inevitably in humans who occasionally err. Employing fault-tolerance techniques becomes imperative to shield end users from specific fault types.

Scalability, on the other hand, necessitates the formulation of strategies to maintain optimal performance, particularly when facing heightened loads. To delve into scalability, it becomes essential to establish quantitative methods for describing load and performance. An illustrative example is Twitter's home timelines, which serve as a depiction of load, and response time percentiles providing a metric for measuring performance. In a scalable system, the ability to augment processing capacity becomes pivotal to sustaining reliability amidst elevated loads.

The facet-rich concept of maintainability essentially revolves around enhancing the working experience for engineering and operations teams interacting with the system. Thoughtfully crafted abstractions play a crucial role in mitigating complexity, rendering the system more adaptable and modifiable for emerging use cases. Effective operability, characterized by comprehensive insights into the system's health and adept management methods, also contributes to maintainability.

Regrettably, there exists no panacea for achieving instant reliability, scalability, or maintainability in applications. Nonetheless, discernible patterns and recurring techniques emerge across diverse application types, offering valuable insights into enhancing these critical attributes.

Data Transformation:

Data Transformation is a vital part of the data flow process, where data changes to meet specific requirements of the process or system. There are two main directions for implementing data transformation: batch processing and real-time processing.

In batch processing mode, data is processed in batches, often scheduled for processing at predefined intervals. This is suitable for tasks requiring the processing of large and complex datasets without an immediate response.

On the contrary, real-time processing is the method of processing data as soon as it arrives, without waiting for a large amount of data to accumulate. This is often preferred in applications demanding low latency, such as real-time event processing.

ETL:

ETL, an acronym for Extract, Transform, Load, stands as a fundamental methodology indispensable for orchestrating the seamless movement of data within storage ecosystems. This three-step process plays a pivotal role in shaping the lifecycle of data.

Firstly, in the "Extract" phase, data is sourced from diverse origins, ranging from databases to files and online services. This initial step lays the groundwork by retrieving relevant information from the varied reservoirs of data.

Subsequently, in the "Transform" phase, the extracted data undergoes a metamorphosis to align with the specific requirements of the target system. This transformative stage encompasses tasks such as data cleansing, format conversions, and even the computation of novel indices, ensuring the data is refined and tailored to suit its intended purpose.

Finally, the "Load" phase marks the culmination of the ETL process. At this juncture, the meticulously transformed data finds its destination, being loaded into the designated storage system. This storage system typically takes the form of a data warehouse or data lake, serving as the repository for the refined and purpose-adapted data. In essence, ETL encapsulates a

systematic and indispensable approach to managing the intricate journey of data within the expansive realm of storage systems.

Data Pipe:

A data pipe is an essential concept in deploying effective data flow. Built on the idea of automating the movement and transformation of data, data pipes serve as powerful workflow streams.

Through the use of data pipes, large data volumes can be processed flexibly and efficiently. Tasks such as error handling, performance monitoring, and even deploying new transformations can be automated, minimizing manual intervention and enhancing system stability.

Data Orchestration:

Data Orchestration is the intricate process of coordinating and managing multiple data processes, workflows, or services to achieve specific outcomes. At its core, it involves the meticulous definition and management of workflows, determining the sequential order, and dependencies among various data processes. Task execution is finely tuned through controllers that schedule and orchestrate tasks at optimal times and in a precise sequence, ensuring the desired outcomes are achieved. Controllers play a pivotal role in managing dependencies between tasks, orchestrating the execution of tasks only when their dependent tasks are successfully met. Robust orchestration systems provide comprehensive tools for monitoring workflow progress and logging pertinent information, offering insights to address and rectify any potential issues. Moreover, controllers optimize performance by strategically breaking down complex tasks into multiple subtasks, executing them in parallel to enhance overall system efficiency.

2.2 Basic of meteorology

Meteorology is the scientific study of the atmosphere that focuses on weather processes and forecasting. It involves the study of the behavior, dynamics, and physical properties of the earth's atmosphere and how these factors affect the weather and climate. Meteorologists use various tools and techniques, including weather satellites, radars, and computer models, to predict weather conditions and to understand the complex interactions within the atmosphere.

Meteorology is important for several reasons, such as predicting weather to warn of severe conditions, understanding climate change, and aiding in decision-making for agriculture, aviation, and other industries dependent on weather.

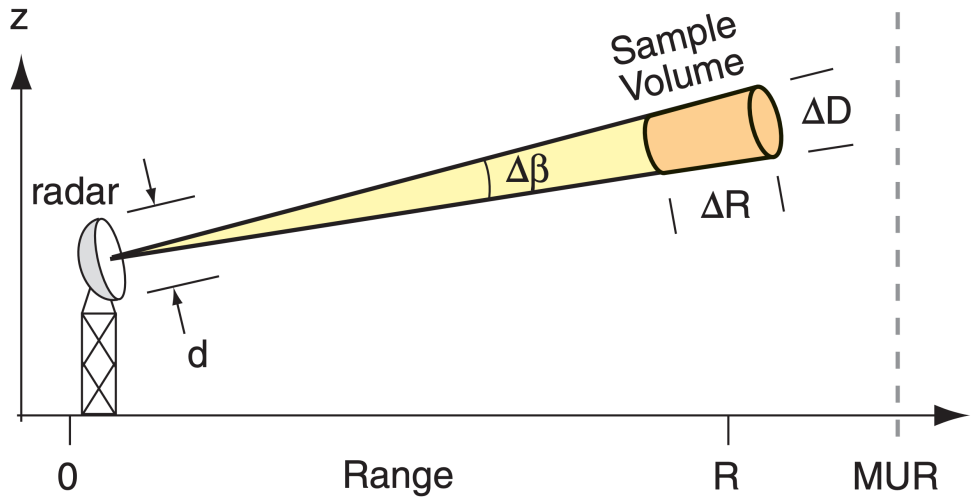


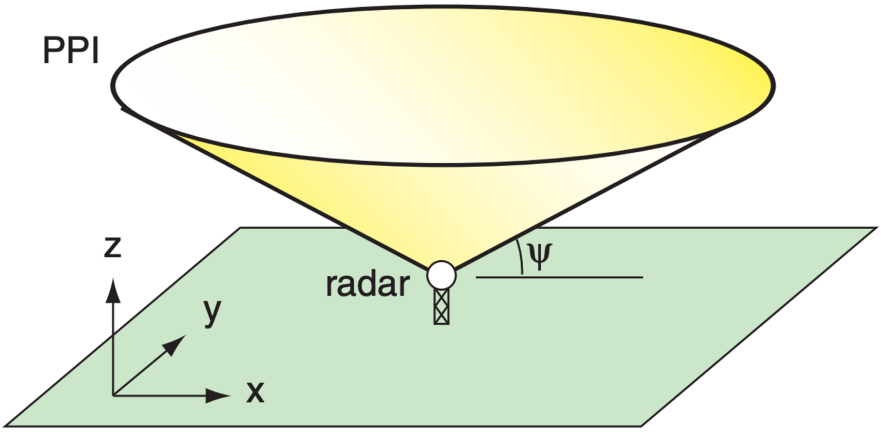
Figure 2.1: *A typical meteorological radar [17]*

2.2.1 Basic terminologies

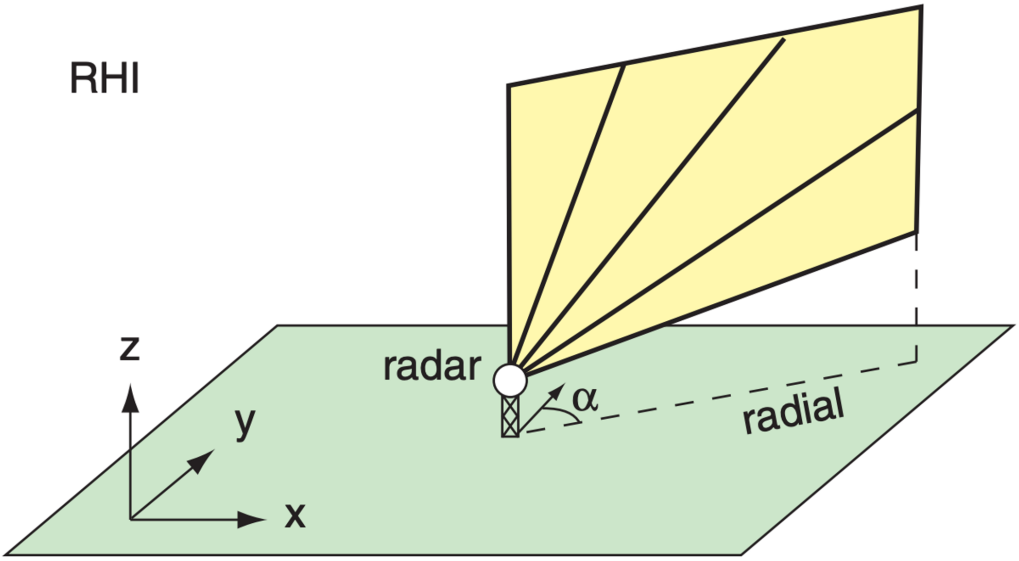
Weather Radar

Weather radar, short for weather surveillance radar, is a type of radar system used to detect and monitor precipitation, as well as other atmospheric phenomena such as the movement of severe weather systems. It plays a crucial role in meteorology and helps meteorologists track and forecast weather conditions. Normally, weather radars are programmed to scan in an azimuth of 360^o. For every round, the radar will scan at a different altitude. It usually takes about four to ten minutes for the radar to complete a full scan.

For PPI representation, the radar will scan the entire azimuth, but only at a certain altitude. The final result would be similar to a map on a flat surface. For RHI, in contrast, the radar retains the azimuth but increases in altitude. The collected result gives viewers more details about the height and sizes of a meteorologist event.



(a) *Plan-Position Indicator - PPI - [17]*



(b) *Range Height Indicator - [17]*

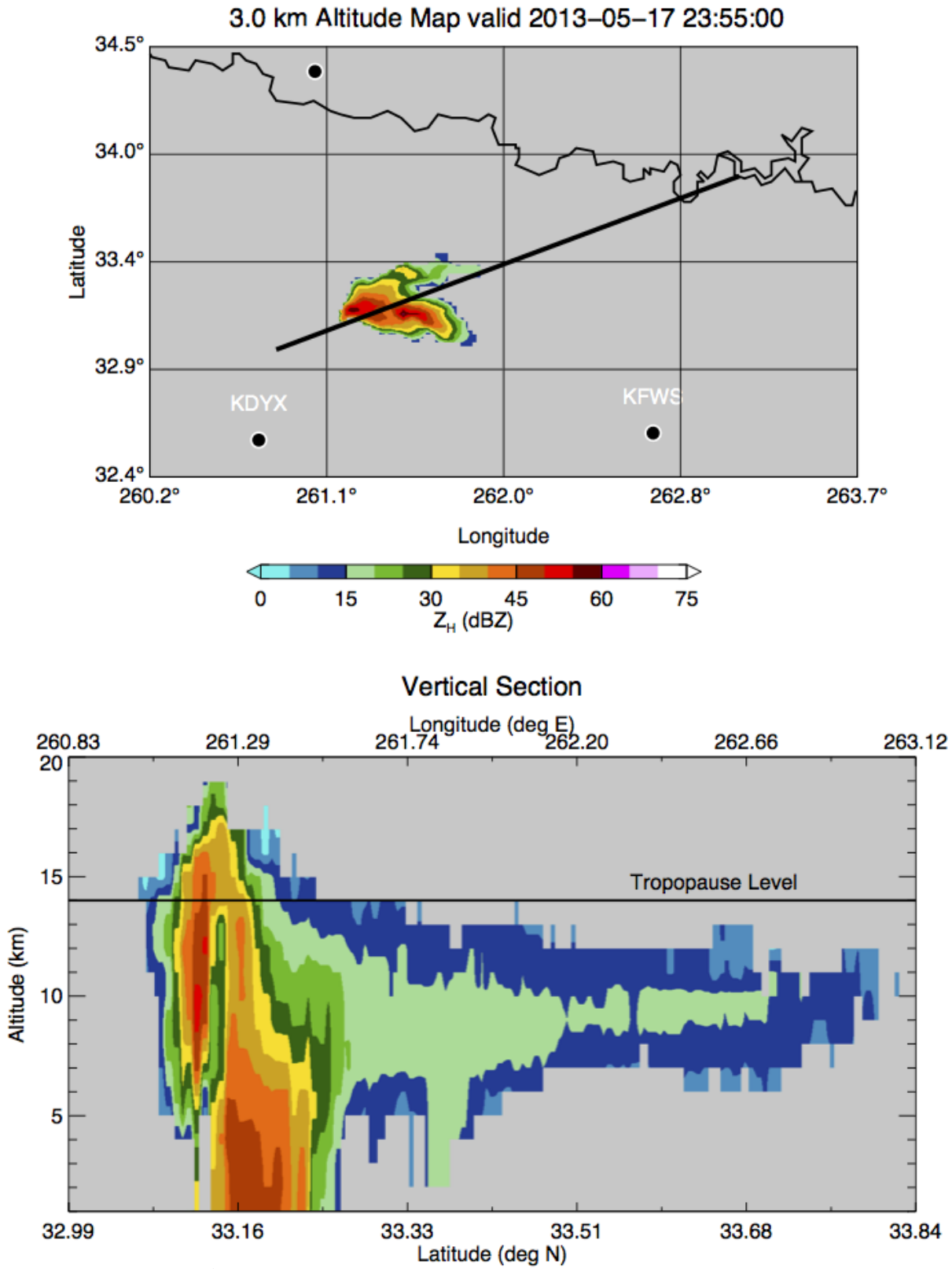


Figure 2.3: *Comparing the result between PPI and RHI - [2]*

Radar equation and Reflectivity

At a certain point in time, weather radar will emit a short pulse of radio wave ($\Delta t = 0.5 - 10\mu s$). Depending on the density of free molecules in the air (water vapor, smoke, ...), the energy of this wavelength will be partially absorbed. The wavelength intensity that the radar receives will be less than the intensity of the original wave. This ratio is expressed through **The radar equation** [17]:

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

Which, the variables of the equation include:

- $|K|$ unitless:
  - $|K|^2 \approx 0.93$ for droplets
  - $|K|^2 \approx 0.208$ for ice crystal
- $R(\text{km})$: distance from the radar to the target
- $R_1 = \sqrt{Z_1 \cdot c \cdot \Delta t / \lambda^2}$: ratio of distance
- Z: Radarś reflectivity
- $Z_1 = 1\ \text{mm}^6\ \text{m}^{-3}$: Radar's unit reflectivity

From the radar equation, we can derive the formula for reflectivity:

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

Meteorologists are usually interested in this number because it is proportional to the amount of precipitation.

Value (dBZ)	Weather
-28	Haze
-12	Clear air
25 - 30	Dry snow / light rain
40 - 50	Heavy rain
75	Giant hail

Table 2.1: *Relation between reflectivity and precipitation - Stull [17]*

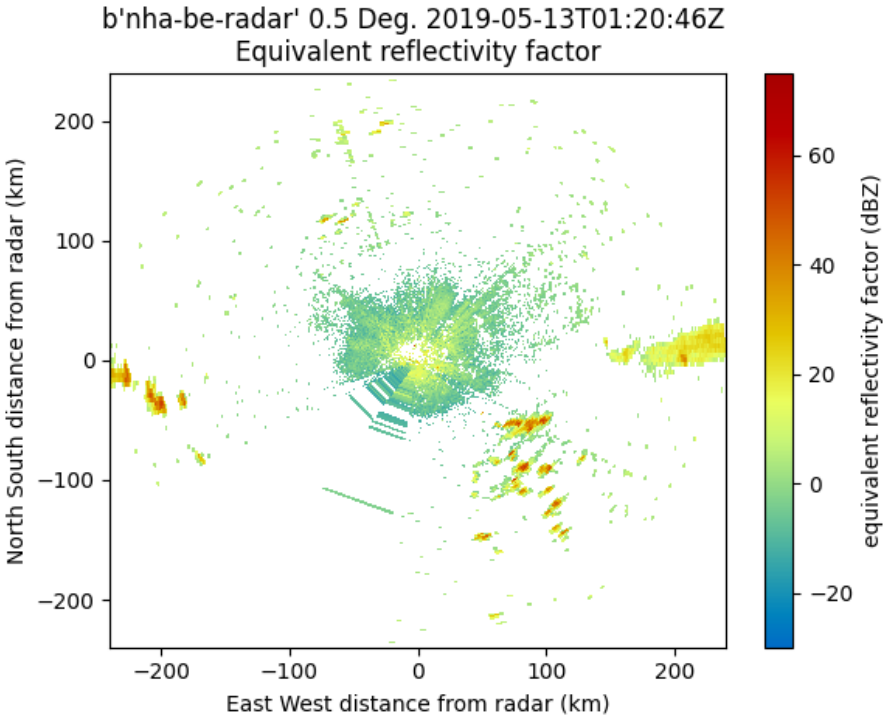


Figure 2.4: *Reflectivity from Nha Be radar*

Radial velocity

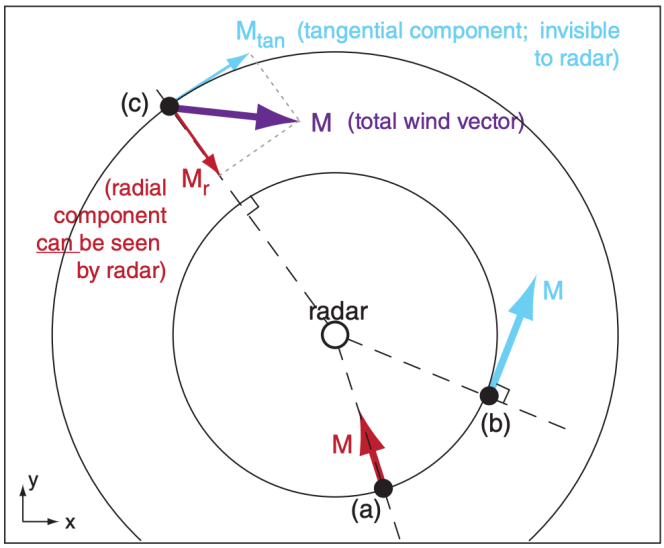


Figure 2.5: *Illustration for the velocity situations that a Doppler radar can observe. (a) When the wind direction at point M coincides with the radius of the circle centered at the radar, the radar can determine the velocity at this point. (b) When the wind direction is tangent to the circle, the radar cannot determine the velocity. (c) Analyzing the wind direction at M into two perpendicular velocities, the radar can only determine the velocity vector along M_r. - Stull [17]*

When the radio waves from these Doppler radars propagate to the molecules in the air, the displacement of these particles causes a phase shift between the transmitted and received signals. Radars rely on this information to calculate the wind velocity at various points in space.

2.2.2 Radar Products

Echo Top Height (ETH)

In radar meteorology, an echo top signifies the highest altitude at which precipitation particles are detected. Essentially, it pinpoints the maximum elevation angle where the intensity of precipitation, quantified by reflectivity, surpasses a predetermined threshold. This information is crucial for understanding the vertical extent of precipitation systems and their potential impact.

The modified ETH algorithm, developed by Lakshmanan et al. (2013) [9], can be applied to a NEXRAD PPI volume scan through a systematic approach. Initially, the maximum elevation angle, denoted by θ_b, where the reflectivity, Z_b, exceeds a predefined echo-top reflectivity threshold, Z_T (e.g., 0 dBZ, 18 dBZ), is located. If θ_b is not the highest elevation scan available

in the volume, the reflectivity value, Z_a, at the subsequent higher elevation angle, θ_a, is obtained. The echo-top height, θ_T, is then calculated using the following equation:

$$\theta_T = \frac{(Z_T - Z_a)(\theta_b - \theta_a)}{Z_b - Z_a} + \theta_b$$

In cases where θ_b coincides with the highest elevation scan accessible, θ_T is set to $\theta_b + \beta/2$, where β represents the half-power beamwidth. This scenario typically occurs either far from the radar, where higher elevation scans have shorter ranges compared to a baseline "surveillance" scan, or very close to the radar, where the highest elevation scan fails to capture the cloud's peak.

Under these circumstances, θ_T corresponds to the top of the beam containing reflectivity greater than or equal to Z_T. Thus, the traditional echo-top computation is adopted when no data exists from higher elevation scans.

The mathematical formulation of this can be expressed as:

$$\theta_T = \begin{cases} \frac{(Z_T - Z_a)(\theta_b - \theta_a)}{Z_b - Z_a} + \theta_b & \text{if } \theta_b \text{ is not the highest elevation scan} \\ \theta_b + \frac{\beta}{2} & \text{if } \theta_b \text{ is the highest elevation scan} \end{cases}$$

This equation incorporates a conditional statement to account for the two scenarios based on the availability of data from higher elevation scans.

Plan Position Indicator (PPI)

The Plan Position Indicator (PPI) is the most common type of radar display. It shows a horizontal slice of radar data at a fixed elevation angle. The PPI scan provides the intensity of returned echoes from targets such as raindrops, snowflakes, or hail. The radar emits pulses at a specific elevation angle θ and records the returned signal. The range r to the target is calculated based on the time delay of the returned signal:

$$r = \frac{c\Delta t}{2}$$

where c is the speed of light and Δt is the time delay. The PPI displays this data in polar coordinates, with the azimuth angle ϕ and range r as the primary variables. This display helps in determining the structure of storms, the intensity of precipitation, and features such as fronts

and squall lines.

Constant Altitude Plan Position Indicator (CAPPI)

The Constant Altitude Plan Position Indicator (CAPPI) provides radar reflectivity data at a constant altitude above the radar site. This product is created by interpolating data from multiple elevation scans to represent reflectivity at a specified altitude h. The interpolation uses the reflectivity values $Z(\theta_i, r_i)$ from different elevation angles θ_i and ranges r_i, and projects them to the desired constant altitude:

$$Z_{CAPPI}(h) = f(Z(\theta_1, r_1), Z(\theta_2, r_2), \ldots, Z(\theta_n, r_n))$$

where f denotes the interpolation function. CAPPI overcomes limitations posed by the Earth's curvature and terrain obstructions, providing a clearer picture of precipitation distribution at a constant altitude, aiding in tracking and analyzing the horizontal spread and intensity of precipitation within storms.

Maximum Echo Height (HMAX)

Maximum Echo Height (HMAX) displays the maximum altitude at which significant radar echoes are detected within each radar sweep area. HMAX is calculated by determining the highest elevation angle θ_i at which the reflectivity Z_i exceeds a predefined threshold Z_T. The height h_i corresponding to θ_i and range r_i is given by:

$$h_i = r_i \sin(\theta_i) + H_r$$

where H_r is the radar height above sea level. The maximum height H_{max} is then:

$$H_{max} = \max(h_i \mid Z_i \geq Z_T)$$

HMAX helps in understanding the vertical development of thunderstorms and is used to gauge storm intensity and potential severe weather occurrences.

Composite Maximum Reflectivity (CMAX)

Composite Maximum Reflectivity (CMAX) provides a top-down view of the highest reflectivity measured at any altitude within the volume of a radar scan. CMAX is generated by compiling the maximum reflectivities Z_{max} from all elevation scans into a single image. For each horizontal grid point (x, y), CMAX is given by:

$$Z_{CMAX}(x, y) = \max(Z(\theta_i, r_i) \mid (x_i, y_i) \in (x, y))$$

where (x_i, y_i) are the horizontal coordinates corresponding to the radar coordinates (θ_i, r_i). CMAX is valuable for identifying the strongest areas of storms across multiple altitudes, highlighting potential regions of severe weather. It allows forecasters to quickly assess the overall intensity of precipitation and identify the most significant storm cells over a large area.

2.3 Common Data format in meteorology

2.3.1 SIGMET data format - raw format (Vaisala)

Vaisala, a Finnish company renowned for its expertise in environmental and meteorological instrumentation, has developed the RAW format, also known as SIGMET [11], as a specialized storage format for organizing data output from its radar devices. This format represents a meticulously designed system tailored to efficiently manage the voluminous data generated by radar scanning sessions.

Distinctive features characterize the architecture of the RAW format. Notably, the file content is partitioned into discrete blocks, each precisely sized at 6144 bytes. This particular sizing aligns with historical conventions rooted in the main storage capacities of older tape devices, reflecting a pragmatic approach to data organization and storage optimization. Furthermore, the RAW format typically consolidates data from multiple radar scanning sessions within a single file, fostering data aggregation and accessibility.

Within each block of the RAW format, data records are meticulously structured, ensuring efficient utilization of storage space. In instances where space within a block remains unoccupied after accommodating data records, padding with additional zeros is employed, maintaining structural integrity and facilitating streamlined data retrieval processes.

The adoption of the RAW format confers several notable advantages, as elucidated by empirical analyses and industry insights [18]. Firstly, the format demonstrates inherent compatibility with a diverse array of tape types, a legacy consideration owing to the historical prevalence of tape-based storage systems. Despite the evolution of storage technologies, tapes persist as cost-effective alternatives, underscoring the ongoing relevance and practicality of the RAW format in contemporary data management contexts. Moreover, the block-oriented architecture of the RAW format facilitates robust error recovery mechanisms at the storage system level, enhancing data reliability and resilience against potential hardware failures.

Despite its merits, concerns persist regarding the alignment of the storage structure delineated by the RAW format with the corresponding mappings on hard drives and tape devices. Addressing these concerns necessitates meticulous attention to compatibility and interoperability considerations, ensuring seamless data interchangeability across heterogeneous storage environments.

In essence, the RAW format epitomizes Vaisala's commitment to innovation and excellence in the domain of environmental instrumentation. Its tailored design and inherent advantages render it a formidable tool for efficiently managing and organizing radar data, thereby empowering researchers and practitioners in meteorology and related fields with robust data management capabilities.

2.3.2 NETCDF data format - Network Common Data Form

The Network Common Data Form (NetCDF) [14] represents a pivotal file format meticulously engineered for the purpose of accommodating multidimensional scientific data. Embedded within the framework of the NetCDF library system are a variety of binary formats, each strategically contributing to the overarching flexibility and scalability of data management within scientific domains. These formats, which have evolved over iterations of the NetCDF library system, are emblematic of its adaptability to diverse data requirements and computational environments.

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

Figure 2.6: *Radar information in NETCDF format. The total dimensions of the dataset are 2975, grouped into 4 distinct labels.*

Foremost among these formats is the Classic Format, which has been a mainstay since the inception of the NetCDF library and continues to serve as the default option for file creation. Over time, subsequent releases of the NetCDF library system have introduced additional formats, each with distinct features tailored to specific demands. The 64-bit Offset Format, for instance, emerged with version 3.6.0, addressing the need for expanded variable and file sizes beyond the limitations of its predecessors.

In a similar vein, the integration of the netCDF-4/HDF5 Format with the release of version 4.0 represents a milestone in the evolution of the NetCDF ecosystem. By harnessing the capabilities of the Hierarchical Data Format version 5 (HDF5), this format bridges the gap between the NetCDF and HDF5 data models, offering users enhanced capabilities such as support for unlimited dimensions and substantially larger file sizes. Furthermore, the incorporation of the HDF4 SD Format and the CDF5 Format underscores the NetCDF ecosystem's commitment to interoperability and compatibility with parallel processing frameworks.

Central to the design philosophy of NetCDF formats is the concept of self-description, wherein comprehensive metadata, including attribute name/value pairs and data array structures, are encapsulated within the file header. This design not only fosters platform independence but also facilitates seamless data exchange and interoperability across diverse computational environments.

To illustrate the versatility and efficacy of the NetCDF format, consider a practical scenario involving the storage of meteorological parameters, including temperature, humidity, pressure, wind speed, and direction, within NetCDF files. This exemplifies the format's capacity to accommodate complex, multidimensional datasets characteristic of scientific research domains,

thereby empowering researchers with a robust and flexible tool for data management and analysis.

Furthermore, it is noteworthy that NetCDF Classic and 64-bit Offset Formats have garnered international recognition as standards endorsed by the Open Geospatial Consortium (OGC), underscoring their role as reliable and interoperable solutions for geospatial data management [13]. This validation not only speaks to the robustness and reliability of the NetCDF format but also underscores its global scalability and compatibility with established standards and practices in the scientific community.

2.4 Related technologies

2.4.1 Apache Airflow™

Apache Airflow™ stands as an open-source platform designed to manage data flow within systems associated with data. In the face of the escalating challenge of data pipeline management, Airflow emerges as a comprehensive solution, automating and optimizing data-related workflows effectively [1].

Airflow not only aids in defining and managing the start and end times of each data pipeline but also provides precise and detailed monitoring of the results of each task. This becomes particularly crucial when ensuring the integrity and reliability of the processed data.

With the ability to discern complex relationships between tasks through the Directed Acyclic Graph (DAG) model, Airflow empowers administrators with tighter control and flexibility in handling workflow processes. Its robust integration with logging systems facilitates detailed activity tracking, assisting in issue resolution and ensuring that every process aligns with expectations.

Simultaneously, the scheduling flexibility makes Airflow an excellent tool for time and resource management. Its strong integration with various data sources and extensibility through plugins allows Airflow to meet diverse needs in data processing and task automation.

Apache Airflow not only delivers robust performance but also brings flexibility and optimal technical features to data processing workflows. With its time management capabilities, powerful logging integration, scheduling flexibility, and scalability, Airflow stands as the top choice for enhancing performance and control in data processing workflows.

2.4.2 Kubernetes

Kubernetes, an open-source system for managing and deploying highly flexible applications in cloud and data center environments, has evolved into one of the most widely adopted tools in the field of Information Technology [8]. Originally developed by Google and later transferred to the Cloud Native Computing Foundation (CNCF), Kubernetes aims to automate the deployment, scaling, and management of containerized applications, alleviating the burden on developers and system administrators. The platform offers a unified foundation for deploying, scaling, and managing containerized applications across multiple servers.

Kubernetes operates based on key concepts such as Pods, Services, ReplicaSets, and various other abstractions, creating a flexible environment for application deployment and management. This fosters an environment where developers can easily build applications, and system administrators can efficiently maintain them.

Beyond supporting traditional deployment models, Kubernetes paves the way for innovative strategies like Continuous Deployment (CD) and Microservices. With the ability to automate many aspects of the development and deployment process, Kubernetes plays a crucial role in constructing and sustaining complex, flexible, and scalable systems.

Kubernetes Components

Introducing essential concepts for managing and deploying applications, Kubernetes provides an effective and flexible environment. The main components of Kubernetes include Pod, ReplicaSet, Deployment, and Service.

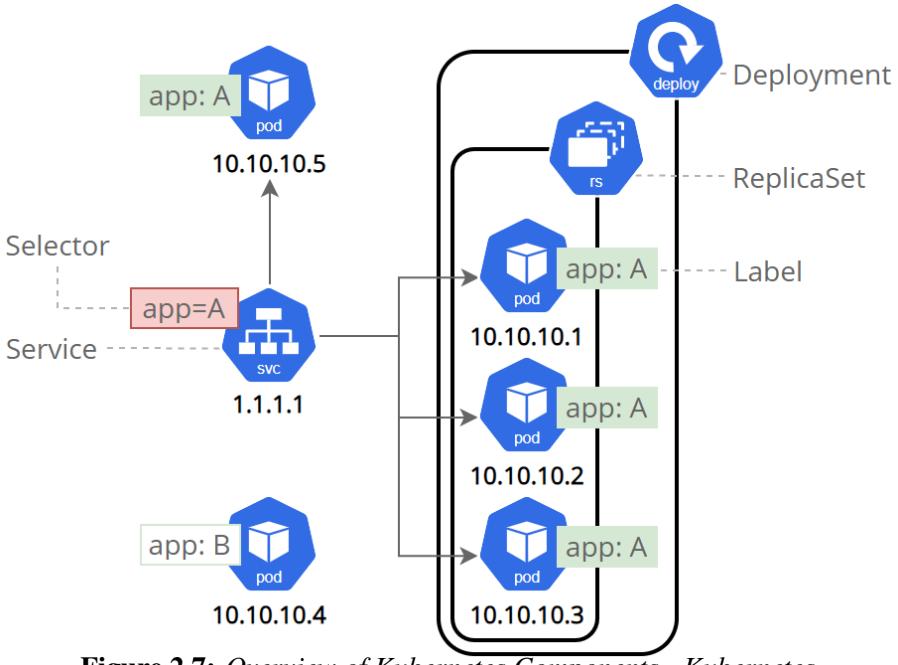


Figure 2.7: *Overview of Kubernetes Components - Kubernetes*

In Kubernetes, a **Pod** serves as the fundamental unit, representing a collection of containers that share a common workspace. Within the same Pod, containers collaborate by sharing network and storage resources, fostering interaction and enabling the construction of intricate applications.

The **ReplicaSet**, a crucial resource in Kubernetes, ensures a designated number of Pods operate in a specified manner. In the event of a Pod failure or shutdown, the ReplicaSet automatically initiates the creation of a new Pod to replace it. This mechanism ensures the application's stable state by guaranteeing a defined number of Pods are consistently operational.

For managing the deployment and updating processes of applications, **Deployment** is a key component in Kubernetes. It articulates the desired state of the application and orchestrates the updating of the ReplicaSet to achieve that state. Deployment provides versatile management capabilities, facilitating the deployment of new versions, rollbacks, and updates without disrupting the service.

The **Service** resource in Kubernetes furnishes an HTTP port to Pods, generating a unique IP address and DNS name for a cluster of Pods. This enables seamless communication among applications within the cluster and with external environments. Service effectively simplifies the intricacies of handling multiple Pods and IP addresses, offering a straightforward means of accessing services within the Kubernetes environment.

Typically, large-scale systems leverage Kubernetes in their software development and de-

ployment processes. This adoption brings several advantages, including efficient resource management and self-recovery capabilities. Kubernetes optimizes resource utilization, ensuring optimal performance and reducing waste. Additionally, it automatically addresses issues during operations, enhancing high availability.

However, the technology is not without its challenges, including a steep learning curve for beginners. Mastery of diverse knowledge areas such as computer networking and containerization is necessary. Moreover, deploying and maintaining Kubernetes demands significant resources, both in terms of personnel and hardware, particularly for smaller organizations.

High Availability in Kubernetes

High Availability is a crucial factor in the success of any system. In Kubernetes, High Availability is achieved through the combination of several features, including self-healing, load balancing, and auto-scaling.

First, Kubernetes uses **Deployment**, a type of **Controller** for managing replicas. Through Deployment, we can easily perform horizontal scaling, which is the process of increasing the number of replicas of a Pod. This mechanism ensures that the application can handle numerous requests without compromising performance. Moreover, in case of a Pod failure, a Deployment makes sure that a new Pod is created to replace it, ensuring the amounts of predefined replicas is always maintained.

At network layers, Kubernetes uses **Service** for communication between various components within or outside the cluster. Instead of directly accessing Pods, other components can access Services, which will redirect the request to the appropriate Pod. When Pods are replaced, the Service will automatically update the routing rules to ensure the request is sent to the correct Pod. Outside the cluster, Kubernetes also uses **Ingress** to manage external access to Services. Instead of specifying the direct node IP address, Clients can abstract it by using only the URL or hostname.

Finally, at the storage level, Kubernetes provides Persistent Volumes. By doing so, applications can be agnostic to the underlying storage infrastructure. This allows for easier management and scaling of storage resources. Not only that, depending on the provided **Storage Class**, Kubernetes makes sure that the data is replicated to multiple nodes, ensuring data availability in case of node failure.

To summarize, Kubernetes provides a robust set of features to ensure High Availability. By leveraging these features, we can build a highly available system that can scale with our load, while also being resilient to failures.

2.4.3 LROSE

LROSE (Lidar Radar Open Software Environment) is a project supported by the National Science Foundation (NSF) with the goal of developing common software for the Lidar, Radar, and Profiler community. The project operates based on the principles of collaboration and open source. The core software package of LROSE is a collaborative effort between Colorado State University (CSU) and the Earth Observing Laboratory (EOL) at the National Center for Atmospheric Research (NCAR) [7].

Originating from the need for a unified software environment for processing Lidar and Radar data in atmospheric science research [7], the project addresses complexities related to integrating data from various observation platforms, including Lidar, Radar, and Profiler. These components are designed to meet the specific needs of meteorologists and researchers working with remote sensor data.

LROSE is widely used in meteorological research, including studies related to cloud and precipitation processes, boundary layer dynamics, and other meteorological phenomena. The software supports the analysis of observation data collected from ground-based tools such as Lidar and Radar. LROSE seamlessly integrates with a variety of model and atmospheric analysis tools to optimize its capabilities. Researchers often integrate LROSE into numerical weather prediction models as well as other data assimilation techniques, creating a flexible and powerful system.

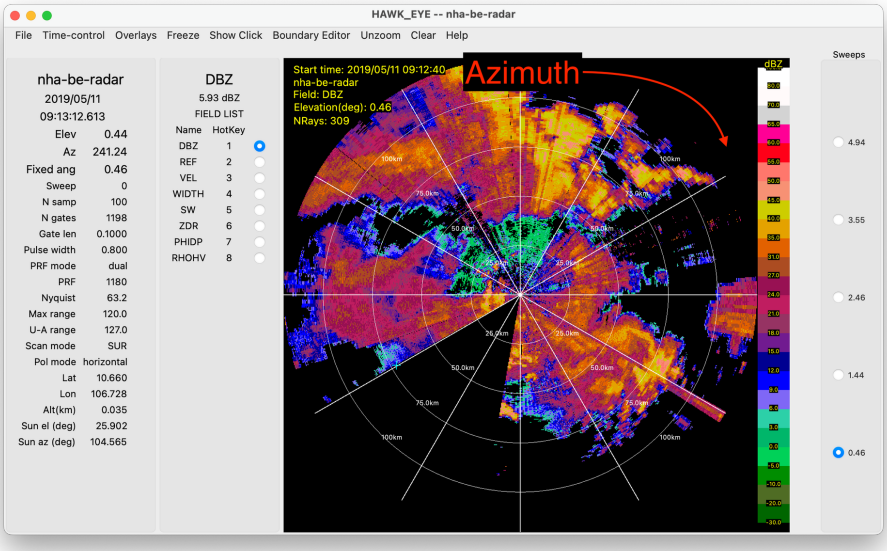


Figure 2.8: *Hawk Eye, Lidar and Radar visualization tool of LROSE*

The project actively encourages participation from a large scientific community, promoting the exchange of ideas, algorithms, and improvements for the software. Regular updates and contributions from users contribute to the continuous development and refinement of LROSE.

2.4.4 Py-ART

In the realm of meteorology, weather radar plays a crucial role in understanding and forecasting precipitation, cloud cover, and other atmospheric phenomena. Extracting meaningful information from this data requires specialized tools and techniques. Py-ART (the Python ARM Radar Toolkit) is an open-source Python library designed specifically for working with weather radar data.

Developed by the Atmospheric Radiation Measurement (ARM) Climate Research Facility, Py-ART offers a comprehensive suite of algorithms and utilities for researchers and atmospheric scientists. However, its user-friendly design and extensive functionalities make it valuable for anyone working with weather radar data, from meteorologists to students.

Core functionalities

Py-ART transcends its role as a simple data repository. It empowers users with a comprehensive suite of functionalities, transforming raw weather radar data into meaningful insights. At its core, Py-ART offers a robust toolbox for data manipulation, analysis, and visualization,

catering to the diverse needs of meteorologists, researchers, and anyone working with weather radar information.

The journey begins with seamless data access. Py-ART supports a wide range of file formats commonly used in atmospheric research. This versatility allows users to import data from various sources, eliminating compatibility hurdles and streamlining the workflow. Once the data is loaded, Py-ART's processing capabilities come into play. Real-world radar measurements are not without imperfections. Signal attenuation, caused by factors like distance and intervening obstacles, weakens the returning signal. Py-ART provides algorithms to correct for this attenuation, ensuring the accuracy of the retrieved information. Furthermore, raw radar data often contains noise – unwanted electrical disturbances that can distort the signal. Py-ART offers a variety of filtering techniques to remove this noise, resulting in a cleaner and more reliable dataset. Calibration, a crucial step in ensuring data integrity, is also made possible by Py-ART. By applying calibration techniques, users can account for systematic biases inherent in the radar system, leading to more precise measurements.

Once the data is cleaned and processed, Py-ART shines in its ability to visualize and analyze this information. It seamlessly integrates with popular scientific Python libraries like Matplotlib. This powerful combination allows users to create informative and visually compelling representations of the data. Imagine generating high-resolution reflectivity maps that paint a vivid picture of precipitation intensity across a region. Py-ART facilitates this by converting reflectivity data into maps, allowing for easy identification of areas with heavy rain or snowfall. Beyond maps, Py-ART enables the creation of vertical profiles, which depict how reflectivity and other parameters vary with altitude. This provides a detailed understanding of the vertical structure of storms and precipitation events. Analyzing these visualizations in conjunction with environmental data, which Py-ART can also incorporate, allows researchers to delve deeper into the atmospheric processes at play.

Py-ART's functionalities extend beyond basic visualization. It offers advanced analysis tools for tasks like feature detection and storm tracking. Imagine automatically identifying and tracking the movement of severe weather features like hailstorms or tornadoes within radar data. Py-ART's algorithms can accomplish this, providing crucial information for issuing timely weather warnings and protecting lives. Perhaps the most impactful functionality lies in quantitative precipitation estimation (QPE). By analyzing radar data and incorporating environmental factors,

Py-ART can estimate the amount of precipitation that has fallen over a specific area. This information is invaluable for flood forecasting, water resource management, and understanding overall precipitation patterns.

Benefits

Py-ART offers a range of benefits for working with weather radar data, making it a compelling choice for meteorologists and atmospheric scientists alike. Firstly, Py-ART is open-source and freely available, allowing anyone to access, download, and modify its code. This fosters collaboration and innovation within the atmospheric science community, as researchers can contribute improvements and share their work with others.

Additionally, Py-ART boasts cross-platform compatibility, functioning seamlessly on various operating systems including Windows, macOS, and Linux. This ensures wider accessibility and flexibility for users across different computing environments.

The modular design of Py-ART is another key advantage, allowing users to leverage specific functionalities they need for their radar data processing tasks. This modular approach enhances efficiency and adaptability, as users can tailor their workflow to suit their requirements.

Moreover, Py-ART provides extensive documentation and a rich collection of examples, which serve to ease the learning curve for new users. The availability of comprehensive resources empowers users to quickly familiarize themselves with the library's capabilities and effectively utilize its features for their research or operational needs.

2.4.5 Comparing between the tools of LROSE and PyART for Meteorology Data Processing

In meteorology, data processing is essential for interpreting and understanding atmospheric phenomena. While both tools offer robust functionalities for radar and lidar data processing, PyART stands out due to its seamless integration with Python, making it particularly powerful for researchers and developers who leverage Python's extensive scientific ecosystem. This section compares these two tools across several dimensions: functionality, ease of use, extensibility, community support, and performance, with a focus on PyART's advantages.

Functionality

PyART offers a comprehensive suite of tools and features for radar data analysis. It supports data input and output in multiple formats, including NetCDF, HDF5, and MDV, ensuring compatibility with various data sources. In terms of processing algorithms, PyART provides robust options for data correction such as dealiasing and attenuation correction, alongside filtering and moment calculation. For visualization, PyART integrates seamlessly with Matplotlib, enabling users to create detailed 2D and 3D visual representations like PPI (Plan Position Indicator) and RHI (Range Height Indicator) plots. Furthermore, PyART is built on the SciPy ecosystem, which facilitates easy integration with other Python scientific libraries such as NumPy, Pandas, and Scikit-learn, thus significantly enhancing its analytical capabilities.

On the other hand, LROSE-Core is also an extensive suite aimed at standardizing and enhancing radar and lidar data processing capabilities. Developed as an open-source initiative, it emphasizes interoperability and advanced processing techniques. The suite supports a wide range of radar and lidar data formats, including proprietary ones, making it highly versatile. LROSE-Core includes sophisticated signal processing algorithms such as dual-polarization processing, clutter filtering, and Doppler velocity analysis. For visualization, it offers advanced tools through its HawkEye and CIDD applications, which provide interactive and customizable plotting capabilities. Additionally, the toolset is designed to handle real-time data processing, which is crucial for operational radar networks.

Ease of Use - Learning Curve

PyART is designed with simplicity in mind, making it accessible to users with basic programming knowledge. The API is well-documented, and the library includes numerous examples and tutorials. Its seamless integration with Python's scientific stack (NumPy, XArray, and Matplotlib, ...) makes it a natural choice for researchers already familiar with these tools. Python's ease of use and readability further enhance PyART's appeal, allowing users to quickly prototype and deploy analysis workflows.

LROSE-Core, while powerful, has a steeper learning curve compared to PyART. It requires a more in-depth understanding of radar and lidar data processing principles. The installation process can be more complex, especially for users unfamiliar with building software from source. However, the detailed documentation and active community support can help mitigate these

challenges.

Extensibility

PyART's modular design makes it highly extensible. Users can easily incorporate their own algorithms or modify existing ones. Its reliance on Python ensures that it can leverage the extensive range of available libraries for additional functionality. This extensibility makes PyART particularly powerful, as users can integrate machine learning libraries such as TensorFlow or Scikit-learn to develop advanced predictive models directly within their radar data processing workflows.

LROSE-Core is also designed for extensibility, with a focus on providing a comprehensive platform for radar and lidar data processing. Its open-source nature allows users to contribute new algorithms and features. However, extending LROSE-Core may require more specialized knowledge in C++ and radar/lidar processing techniques, potentially limiting its accessibility compared to PyART.

2.4.6 ASP.NET Core

ASP.NET Core is a free, cross-platform, open-source framework for building modern, cloud-based, internet-connected applications. It is particularly well-suited for developing API servers that need to interact with databases, thanks to its robust set of features, high performance, and strong community support.

There are many reasons to believe that ASP.NET and C# can be a good candidate for building server in managing an integrated database system. The following sections explain some key features that Microsoft and this ecosystem provides when constructing the system.

Cross-Platform Support

ASP.NET Core distinguishes itself through its remarkable cross-platform compatibility, heralding a paradigm shift wherein applications developed on this framework can seamlessly traverse the disparate terrains of Windows, Linux, and macOS environments. This intrinsic flexibility underscores the platform's adaptability to diverse deployment scenarios, be it the precincts of a localized server, the ethereal expanse of a cloud platform, or the nuanced amalgam of a hybrid setup. Such versatility empowers developers with the autonomy to cherry-pick

the operating system that resonates most harmoniously with the exigencies of their deployment milieu, thereby engendering a milieu of operational fluidity and infrastructural resilience.

Robust Security Features

Security, an ever-pressing concern in the digital milieu, assumes paramount significance in the realm of web applications, particularly those enmeshed within the labyrinthine landscapes of database interactions. ASP.NET Core, cognizant of this imperative, fortifies its arsenal with a plethora of built-in mechanisms geared towards safeguarding the sanctity of data transactions. It proffers native support for industry-standard authentication protocols and a panoply of security features engineered to combat the insidious machinations of common cyber threats, including but not limited to Cross-Site Scripting (XSS) and Cross-Site Request Forgery (CSRF). Augmenting its robust security posture are provisions for multi-factor authentication and seamless integration with external authentication providers such as Google and Facebook, thereby delineating a formidable bulwark against the incursions of malevolent actors.

Rich Ecosystem and Tooling

At the heart of ASP.NET Core beats the pulse of the expansive .NET ecosystem, a sprawling tapestry interwoven with a myriad of libraries, tools, and frameworks. Foremost among these is Entity Framework Core (EF Core), a venerable Object-Relational Mapper (ORM) that bestows upon developers an intuitive conduit for data access and manipulation. Renowned for its versatility, EF Core extends its embrace to a diverse array of database platforms, including but not limited to SQL Server, SQLite, PostgreSQL, and MySQL, thereby furnishing developers with the wherewithal to navigate the complex terrain of database interactions with consummate ease and finesse.

2.4.7 MinIO

MinIO is an open-source, high-performance, distributed object storage system. It is designed to be fully compatible with the Amazon S3 API, which makes it an attractive option for developers and organizations looking to deploy scalable and cost-effective storage solutions. MinIO's simplicity, speed, and scalability make it suitable for a wide range of applications, from private cloud infrastructures to data lakes and large-scale data processing environments.

S3 Compatibility and Interoperability

A salient hallmark of MinIO lies in its comprehensive compatibility with the Amazon S3 Application Programming Interface (API), a feature that bestows upon it the mantle of interoperability with a plethora of applications tailored for Amazon S3. This alignment with the S3 API translates into a seamless interplay between MinIO and applications engineered for Amazon S3, obviating the need for any modifications or adaptations. Furthermore, this compatibility extends across a gamut of S3 functionalities, including but not limited to bucket policies, multipart uploads, and the generation of presigned URLs, thereby positioning MinIO as a compelling substitute for S3 across manifold scenarios.

High Performance and Scalability

At the heart of MinIO's allure lies its relentless pursuit of high performance and scalability, traits that are quintessential in navigating the exigencies of modern data ecosystems. Engineered with a laser focus on throughput and scalability, MinIO boasts the capability to seamlessly accommodate prodigious volumes of data with unparalleled efficiency. Leveraging a suite of advanced technologies, including erasure coding, bit-rot protection, and data compression, MinIO stands poised to cater to the exacting demands levied by high-stakes domains such as big data analytics and machine learning workloads, positioning itself as a stalwart guardian of performance and reliability.

Security, Flexibility, and Ease of Use

Ensuring the sanctity of data integrity, MinIO espouses a multifaceted approach towards fortifying its security posture. Embracing server-side encryption as a cornerstone, MinIO ensures that data remains impervious to the predations of malevolent actors even when at rest. Moreover, bolstering its security repertoire is the seamless integration of Transport Layer Security (TLS), imbuing data in transit with an impenetrable cloak of encryption. Augmenting its security credentials are robust access control mechanisms, including the granular delineations afforded by Identity and Access Management (IAM) policies, thereby empowering administrators with fine-grained control over data access and manipulation. Furthermore, MinIO's minimalist design ethos renders it eminently facile to install and manage, underscored by an intuitive web-based management console and a potent command-line interface (CLI) replete with

scripting capabilities, thus ensuring a seamless fusion of sophistication and user-friendliness in the realm of storage resource management.

2.5 Study Cases

2.5.1 Neon: A Case Study in Cloud-Native Database Technology

Neon is a PostgreSQL-compatible, cloud-native database that emphasizes scalability and efficiency by separating storage and compute. This architecture allows users to adjust compute resources based on demand, which is particularly useful for applications with variable workloads.

Neon adopts a **serverless** approach where the database scales compute capacity automatically according to the workload. This feature is economically efficient as it ensures that users pay only for the compute resources they actually use, rather than maintaining a constant level of server capacity.

A notable feature of Neon is its capability to create **"branches"** of the database, similar to version control systems for data. This allows developers to test changes in isolation and seamlessly integrate them into the main database as needed, enhancing development flexibility and reducing the risk of errors in the main database.

Neon offers the capability for almost **immediate database snapshots and restores**. This functionality is invaluable for conducting backups or quickly reverting to a prior state in response to issues, thereby enhancing data management and recovery strategies.

By extending PostgreSQL, Neon ensures **compatibility** with existing PostgreSQL applications and tools, facilitating easy migration for businesses without necessitating significant alterations to their current applications.

Neon's serverless model significantly cuts costs by directly aligning resource usage with demand, optimizing both latency and throughput as compute nodes are stateless and dynamically allocated based on workload needs. The architecture's ability to independently scale storage and compute components provides remarkable flexibility and efficiency, making it an ideal solution for handling varying traffic and workloads.

The branching feature positions Neon as an excellent option for environments where development, testing, and deployment are conducted rapidly and in parallel. Moreover, applications

that experience fluctuating traffic, like e-commerce platforms during sales events, stand to benefit greatly from Neon's scalable and cost-effective compute model.

In conclusion, Neon signifies a progressive step in database technology, especially suited for cloud-native applications that demand high flexibility, scalability, and performance. Its compatibility with PostgreSQL ensures broad accessibility for diverse applications, making it a strong candidate for businesses aiming to harness cutting-edge database technology.

2.5.2 Vaisala: Pioneering Precision in Weather and Industrial Measurements

Vaisala is a global leader in weather, environmental, and industrial measurements headquartered in Finland. Founded in 1936, the company has built a reputation for providing comprehensive meteorological, environmental, and industrial measurement solutions. Vaisala's products and services are crucial for ensuring safety in aviation, road traffic, and shipping industries, and for enhancing efficiency in renewable energy projects and various meteorological operations.

Vaisala's innovative approach includes advanced sensors and instruments that measure everything from humidity and temperature to atmospheric pressure, wind speed, and precipitation. These instruments are known for their precision, reliability, and durability, often being used in harsh and demanding environments.Moreover, the company has been operating on a global scale, serving customers in over 150 countries. Its ability to provide accurate and timely data makes it indispensable for weather forecasting services, airports, and maritime operations worldwide. Vaisala's technology also plays a vital role in combating climate change by supporting renewable energy projects and promoting environmental sustainability.

Despite its success, Vaisala faces challenges such as the need for continuous innovation to keep up with rapidly evolving technology and increasing competition in the global market. However, the growing emphasis on environmental issues and renewable energy offers significant opportunities for expansion. Vaisala's commitment to innovation and quality has established it as a leader in the measurement industry. Its dedication to enhancing safety, efficiency, and environmental stewardship continues to drive its success and expansion into new markets.

2.5.3 MeteoSwiss: Safeguarding Switzerland Through Advanced Meteorological Services

MeteoSwiss, as the Federal Office of Meteorology and Climatology, embodies Switzerland's foremost authority in meteorological endeavors, assuming a pivotal role in the realms of weather forecasting, climate monitoring, and the provision of meteorological services paramount to public safety and economic vitality within the nation's borders. At the nexus of its mandate lies the solemn obligation to furnish the Swiss populace with accurate and timely weather information, robust severe weather warnings, and a compendium of long-term climate data, all of which constitute indispensable linchpins in the apparatus of disaster prevention and mitigation, particularly within locales predisposed to the caprices of avalanches, floods, and other meteorologically induced calamities.

Embracing an ethos of technological prowess and innovation, MeteoSwiss harnesses an array of state-of-the-art meteorological instrumentation, ranging from radar systems and automatic weather stations to satellite-based data acquisition methodologies, thereby enabling the incessant surveillance and appraisal of weather conditions across Swiss territories. Moreover, cognizant of the imperatives of global collaboration, the agency actively engages in international partnerships, synergizing efforts to augment weather forecasting accuracy and participating fervently in initiatives aimed at bolstering global climate monitoring endeavors, thus positioning itself as a veritable bastion of meteorological stewardship on the global stage.

However, amidst the mosaic of its triumphs lies the specter of burgeoning challenges, chief among them being the escalating unpredictability of weather patterns, a phenomenon conjectured to be exacerbated by the specter of climate change. Navigating this terrain necessitates an unflagging commitment to perpetual technological evolution and the relentless refinement of predictive models. Furthermore, the agency confronts the imperative of refining its communication strategies to ensure that its clarion calls resonate effectively with the public consciousness, particularly in the crucible of severe weather events wherein timely and cogent dissemination of information stands as a bulwark against catastrophe and upheaval. Thus, MeteoSwiss stands as an indomitable sentinel, safeguarding the welfare and interests of the Swiss populace through its unwavering dedication to meteorological excellence and the ceaseless pursuit of scientific innovation in the service of public good.

Chapter 3

System Analysis and Design

3.1 Exploratory Data Analysis

This section describes the processes and steps that our team took to study the data.

3.1.1 Describe provided data

The provided data resides inside two parent directories:

For the `convert` directory, it seems like the data there has been converted from the other `UF` files. Compared with the original, these files do not store as many necessary fields, with many meteorologist features being left out or not included. For instance, fields like **Reflectivity**, **Mean doppler velocity** and **Doppler spectrum widtah** only appear in the `UF` files, but not in the converted ones.

```
{ 'time': { 'units': 'seconds since 2019-05-13T01:20:07Z', } ... 'fields': {
    'total_power': { 'units': 'dBZ', 'standard_name':
    'equivalent_reflectivity_factor', 'long_name': 'Total power', 'coordinates':
    'elevation azimuth range', 'data': masked_array( data=[ [37.0, 27.5, 37.5,
    ..., --, --, --], [50.0, 41.0, 31.5, ..., --, --, --], [45.0, 41.0, 29.5,
    ..., 0.0, 4.0, 7.0], ..., [35.0, 32.0, 21.0, ..., 3.0, --, --], [33.5, 27.5,
    20.0, ..., --, --, --], [41.0, 31.5, 27.0, ..., --, --, --] ], mask=[
```

```
8     [False, False, False, ...,  True,  True,  True], [False, False,
    False, ...,
9     True,  True,  True], [False, False, False, ..., False, False, False
    ], ...,
10    [False, False, False, ..., False,  True,  True], [False, False,
    False, ...,
11    True,  True,  True], [False, False, False, ...,  True,  True,  True]
    ],
12    fill_value=1e+20, dtype=float32 ), '_FillValue': -9999.0 }, ... } }
```

Listing 3.1: A sample of metadata extracted from UF file

As a result, we decided to use the `UF` files for our study, while temporarily ignoring the `convert` directory.

3.1.2 Comparing PyART and LROSE

From the advice of our professor, at first, we decided to learn about LROSE and try to open the provided data. LROSE is a set of multiple tools and libraries for working with Radar data output. For macOS, support for Homebrew makes it easier to install these tools for local usage. However, for other platforms, downloading and compiling the source code is required, which increases the complexity and time needed to set up the environment.

The reference for LROSE is not very extensive, with many commands requiring special configs and parameters that have not been written down in the documentation. As a result, it took us a lot of time to figure out how to use the tools and libraries. In the end, though, our team was able to open the provided data and visualize it using the LROSE tools, whose name was `HawkEye`.

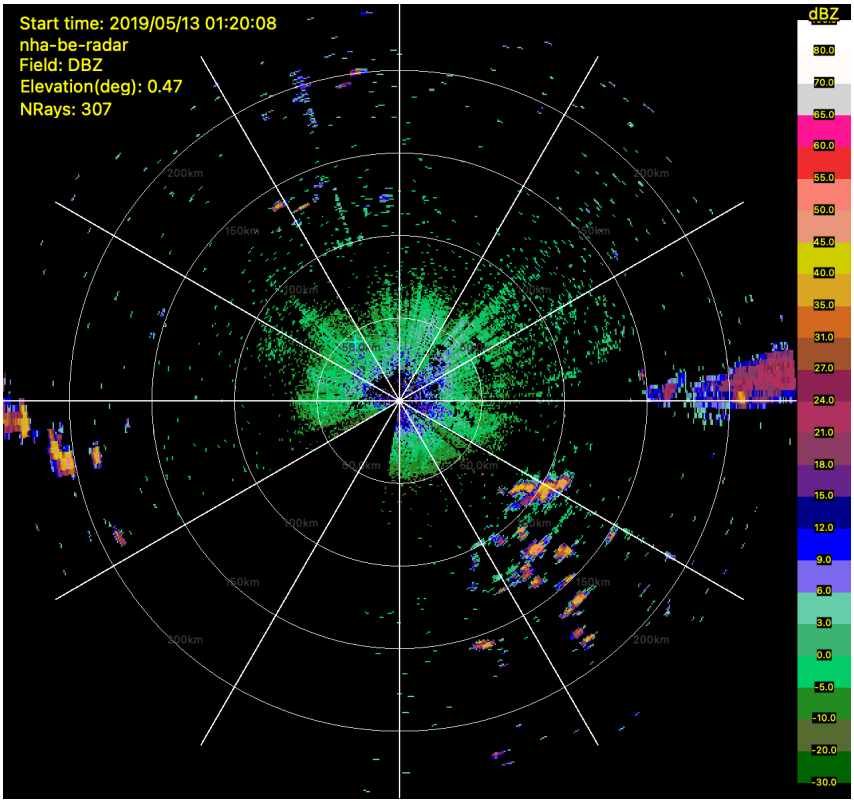


Figure 3.1: *Data visualization of Nha Be radar from HawkEye - Map of Vietnam*

During our investigation, we also encountered a different set of tools that can interact with the radar files, called PyART. When compared with LROSE-core, our team found that PyART contains many more advantages.

First, PyART is a Python library. This means that it is easier to install and use, especially for people who are already familiar with Python. Moreover, by supporting Python, our team can incorporate it into many popular ETL tools that use Python. Take Apache Airflow as an example. Instead of having to write a custom operator to call LROSE-core from the terminal, we can use a Task to perform the same action with PyART. Doing so would reduce much of the complexity of the workflow.

Secondly, with default configurations, PyART represents the data as an XArray [6] format, which is also a popular format for multidimensional data. As XArray is built on top of NumPy [5], this popular library can receive wider adoption when compared with many other formats. Integration with the existing library also means that we can use many other tools to work with the data. For instance, we can use Matplotlib to visualize the data; or use the library SciPy [19] to perform scientific calculations on the data. Because those tools require NumPy, we can use them directly with PyART without having to convert the data to another format.

Finally, PyART also has more comprehensive documentation when compared with LROSE. As newcomers to the subject of meteorology, we found that the documentation of PyART is much more accessible than LROSE. The documentation includes many examples and tutorials. Moreover, the documentation also contains many references to other resources, such as the book *Radar for Meteorological and Atmospheric Observations* [17]. Even though this is a subjective opinion, we believe that the documentation of PyART is much more helpful for further development.

With all the above ideas, our teams decided to use PyART for both the study and the development of the project.

3.1.3 Visualizing data using PyART

When installed, PyART also downloads another Python tool named Cartopy [12]. This library provides many useful features for our understanding of the data, such as the ability to plot the data on a map, or to convert the coordinates from one system to another.

Using these libraries, we can visualize our data directly from a Jupyter Notebook, similar to any other Python data visualization task. The area that the radar collects is a square, with the radar at the center. All of its length extends to about 300 kilometers in each direction.

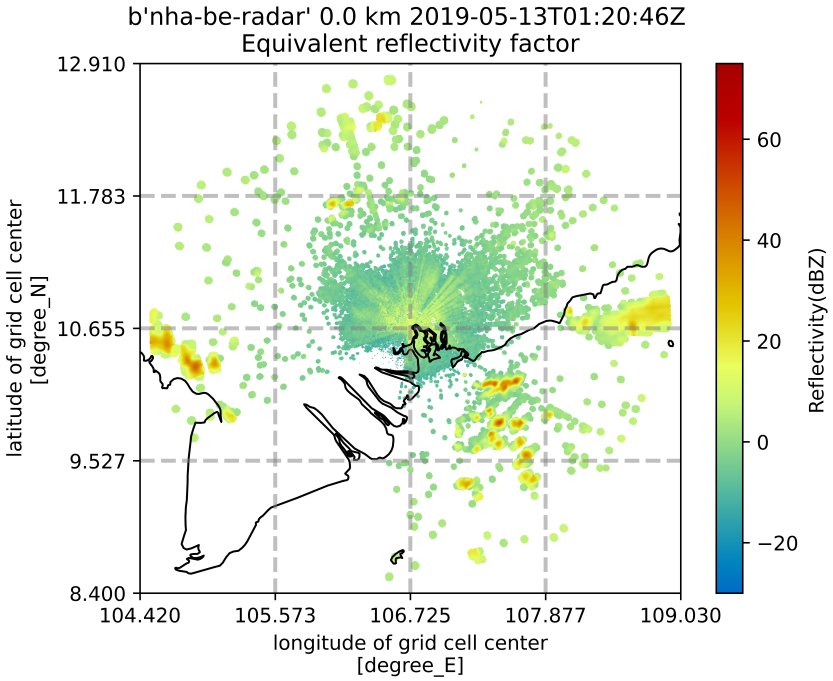


Figure 3.2: *Data visualization of Nha Be radar from PyART - Map of Vietnam - similar zoom to Figure 3.1*

3.2 Requirements

3.2.1 The Need for Business Requirements in Nowcasting

In the realm of meteorology, the importance of accurate and timely weather forecasts cannot be overstated, especially for short-term (3-6 hours) and immediate forecasts known as nowcasting (<3 hours). While automated weather stations are equipped with advanced technology for meteorological data processing—such as data collection, ETL (Extract, Transform, Load) processes, and analysis—the business processes surrounding these forecasts remain largely manual. This disparity between technological capability and operational efficiency hampers the overall effectiveness of weather prediction and response.

Current State of Technology in Meteorological Processing

Automated weather stations have revolutionized the field of meteorology with their ability to continuously monitor and process vast amounts of weather data. These stations employ sophisticated sensors and data processing algorithms to collect, transform, and analyze meteorological data in real-time. The integration of these technologies enables accurate and rapid weather observations, which are critical for creating reliable forecasts.

Challenges in Business Processes

Despite the technological advancements in data processing, many business processes in meteorological services are still conducted manually. This includes constructing reports for abnormal weather phenomena using tools like Microsoft Word and sending email alerts manually. Such manual processes introduce significant delays in issuing alerts and hinder the responsiveness of nowcasting systems. The delay in alert dissemination not only affects the ability to provide timely warnings but also impacts the overall efficiency of weather forecasting operations.

The Need for Automation and Database Handling

To bridge the gap between technological capabilities and operational efficiency, there is a pressing need to automate business processes in meteorological services. An automated system

for handling these processes would significantly enhance the speed and accuracy of weather alerts and nowcasting.

The system should support **dynamic fields for data input**, allowing flexibility in capturing various types of weather data and abnormal phenomena. This dynamic data collection capability ensures that the system can adapt to the diverse and evolving nature of meteorological observations, facilitating the accurate recording of critical data points.

Incorporating **templates for report generation** can streamline the creation of consistent and standardized reports, reducing the time and effort required for manual report construction. Template-based reporting ensures that essential information is presented uniformly, making it easier to interpret and act upon the data provided.

Automating the process of report creation and email distribution ensures that alerts are sent out promptly and accurately, minimizing human error and delays. Automated report creation and distribution systems can trigger alerts based on predefined criteria, ensuring that relevant stakeholders receive timely notifications of abnormal weather conditions.

A **robust database system** is essential to handle the storage, retrieval, and management of data conducted in meteorological business process. This database should provide a reliable platform for data-driven decision-making. Effective database management enables efficient data organization, ensuring that critical information is readily accessible for analysis and reporting.

In conclusion, the integration of automated processes and advanced database management systems is crucial for enhancing the efficiency and effectiveness of nowcasting. By addressing the current limitations in business processes, meteorological services can significantly improve their ability to provide timely and accurate weather forecasts, ultimately safeguarding lives and property.

3.2.2 Functional Requirement

In the initial phase of the data processing workflow, **data ingestion** emerges as a pivotal undertaking, wherein data is systematically amassed from a spectrum of heterogeneous sources, spanning from **Application Programming Interfaces (APIs)** to repositories housing **government data**. This foundational process assumes paramount significance as it sets the groundwork for the acquisition of an extensive array of datasets essential for comprehensive analytical endeavors.

Following data ingestion, the subsequent stage revolves around **data storage** and **management**. This facet encompasses the orchestration of repositories, potentially ranging from conventional databases to more contemporary constructs like **data lakes**. Such infrastructural modalities afford robust frameworks that adeptly cater to the exigencies imposed by the management of copious volumes of data, thereby ensuring efficiency and efficacy in handling data resources.

Concomitant with data management, **data analysis** and **visualization** procedures are meticulously tailored to cater to the discerning requirements of stakeholders entrenched within the meteorological domain and academic research milieu. These analytical tools serve as indispensable instruments, facilitating granular scrutiny and cogent interpretation of meteorological datasets, thereby fostering informed decision-making and scholarly inquiry.

Facilitating seamless interoperability with external platforms and augmenting the system's functionality through synergistic engagement with **APIs** represent pivotal components of the overarching operational paradigm. Such integrative endeavors engender an ecosystem conducive to harmonious interactions with an array of technological adjuncts, thereby amplifying the system's utility and efficacy.

Furthermore, the system incorporates robust **user management** protocols and **access control mechanisms** to ensure the sanctity of data assets, conferring access exclusively to duly authorized personnel. By meticulously safeguarding data integrity and confidentiality, these provisions engender an environment predicated upon trust and accountability, thereby fortifying the organizational edifice.

Finally, the culminating facets of **reporting** and **alerting functionalities** emerge as indispensable conduits for the expeditious dissemination of critical insights and information to an array of stakeholders, encompassing end-users, organizational hierarchies, and strategic decision-makers. These functionalities serve as conduits for the expeditious transmission of salient information, thereby empowering stakeholders to navigate dynamic operational landscapes with acumen and alacrity.

3.2.3 Non-Functional Requirements

In the realm of technological platforms, the facet of **Reliability** stands as a cornerstone, necessitating unwavering commitment towards ensuring persistent high availability and steadfast

operational stability. Paramount to this pursuit is the unyielding dedication towards upholding **data integrity**, **security**, and **privacy** as cardinal imperatives, safeguarding both user information and system data against potential vulnerabilities and breaches.

Moreover, the imperative of **Scalability** looms large, compelling the platform to confront the exigencies posed by burgeoning volumes of data. In this regard, an adept embrace of **cloud-native technologies** and the paradigm of **containerization**, exemplified by stalwarts such as Docker and Kubernetes, emerges as indispensable. Such technological modalities proffer the requisite infrastructure agility, facilitating seamless and expedient scaling of both infrastructure and applications alike, thereby ensuring commensurate responsiveness to evolving operational demands.

Further accentuating the discourse, the criterion of **Performance** assumes primacy, necessitating the provision of expeditious data processing and analysis capabilities. This imperative is underscored by the imperatives of real-time decision-making imperatives and the imperative of managing large-scale data operations with finesse and alacrity.

In tandem, the imperative of **Usability** assumes paramount significance, underscoring the imperative of furnishing an interface that is as intuitive as it is user-friendly. A bespoke interface, meticulously tailored to cater to the diverse exigencies of disparate user cohorts, stands as a testament to the platform's commitment towards ensuring unfettered ease of access and usability.

Moreover, the aspect of **Security** mandates robust fortifications, encompassing both data protection protocols and stringent access control mechanisms. This multifaceted approach is calibrated to thwart the machinations of unauthorized access and avert the specter of potential data breaches, thereby undergirding the sanctity of the platform's operational integrity.

Simultaneously, adherence to **Compliance** imperatives emerges as non-negotiable, enjoining scrupulous conformity with a panoply of regulatory stipulations, data privacy statutes, and overarching compliance requisites. Such fidelity towards regulatory benchmarks serves as an indispensable bulwark, ensuring that the platform operates within the hallowed precincts of legal and ethical propriety.

Lastly, expounding upon the contours of **scaling technology and infrastructure**, the platform espouses an ethos predicated upon the tenets of **cloud-native architectures** and containerization technologies. This strategic pivot affords the platform the requisite nimbleness,

facilitated by the judicious employment of auto-scaling capabilities and the elastic resources proffered by cloud platforms. Furthermore, the adoption of a **microservices architecture** and the decoupling of components augur well for bolstering the platform's resilience and flexibility, enabling independent scaling of discrete services or modules, and thus engendering a paradigm of operational dynamism and adaptability.

3.2.4 Data Requirements

In the intricate domain of radar data management, the foundational phase commences with an exhaustive **cleaning process**. This pivotal endeavor is indispensable, serving as the vanguard against the encroachment of **noise** and extraneous data, thereby ensuring the veracity and reliability of the data corpus. The imperative of cleansing radar files assumes paramount significance, poised as it is to ready them for the rigors of more intricate processing and analysis, which constitute linchpins in facilitating accurate meteorological appraisals.

Subsequent to the meticulous cleansing regimen, the radar data traverses a sophisticated trajectory, wherein it becomes ensconced within a nuanced **processing routine**. At this juncture, specialized algorithms assume the mantle, orchestrating a symphony of analyses to distill meaningful patterns and insights from the data trove. This transformative phase is pivotal, transmuting raw data into actionable intelligence, thereby furnishing decision-makers with a bedrock of dependable insights.

Post the crucible of processing, the data finds sanctuary within the confines of **blobs** nestled within an object storage apparatus, typically epitomized by **Minio**. This methodical selection of storage modality is predicated upon its inherent scalability and facile accessibility, attributes that are deemed indispensable in wrangling the prodigious volumes of data bequeathed by radar systems. Moreover, the architecture of blob storage engenders a milieu conducive to expeditious data retrieval and management, adeptly accommodating the dynamic access patterns germane to meteorological exigencies.

In a bid to further fortify the edifice of data handling efficacy, the apparatus embraces the instantiation of a **caching mechanism**. This strategic deployment bestows upon the system the capability to temporarily harbor recently accessed data, thereby effectuating a palpable reduction in retrieval latencies. This feature assumes outsized significance, particularly in scenarios of heightened exigency, such as severe weather events, wherein the expeditious access to data

assumes the mantle of existential import.

The hallowed precincts of processed and stored radar data resonate with a multifaceted utility, traversing a panoply of applications. Meteorologists, ensconced within the annals of weather forecasting and climate modeling, find themselves inexorably tethered to this fount of data-driven insights. Concurrently, the pantheon of emergency response entities stands as ardent beneficiaries, drawing succor from the real-time data streams that undergird effective decision-making in the crucible of weather-related emergencies. Furthermore, the expansive vista of scientific inquiry finds solace in this rich tapestry of data, wielding it as a potent instrument in unraveling the enigmatic contours of climate patterns and atmospheric dynamics.

3.3 Data management

3.3.1 Blob storage

As an intermediary entity situated between the data producers, exemplified by the radar center, and the data consumers, represented by the Data Model and Machine Learning team, our paramount responsibility lies in ensuring that the data management infrastructure is meticulously designed to seamlessly accommodate the diverse and often contrasting requirements set forth by both parties.

The dataset, procured from the esteemed Nha Be radar center, adheres to the rigorous SIGMET format, as delineated in 2.3.1. This format encompasses a plethora of scalar values pertaining to metadata facets such as the radar's nomenclature, the modality of scanning employed, and the temporal stamp denoting the execution of the scan. Concurrently, this data corpus embodies a rich array of multidimensional metrics encapsulating various meteorological phenomena, including but not limited to reflectivity and energy profiles.

Collaborative engagements with the modeling cadre, predominantly comprising members from the distinguished teams led by Tu and Vinh, have endowed us with a nuanced understanding of their exigencies and prerequisites vis-à-vis data structuring paradigms. From the purview of these erudite teams, data representation assumes two principal manifestations. Firstly, there is a pronounced inclination towards encapsulating data in the form of images tailored to specific geographical demarcations within the Vietnamese domain. A quintessential exemplar of this utilization paradigm is elucidated in Figure 3.2, which encapsulates the data instrumental in honing

the multimodal solutions. Alternatively, there exists a proclivity towards direct data consumption in the form of NumPy Tensors, leveraging the expansive spectrum of NumPy-compatible APIs inclusive of those furnished by SciPy and XArray. Although ostensibly more efficient, this mode of data consumption has not garnered widespread adoption within the precincts of our collaborative endeavor.

Armed with an exhaustive comprehension of the requisites delineated by both stakeholders and guided by heuristic principles germane to the architecture of rudimentary data warehousing systems, we proffer a set of salient properties that our envisaged system ought to embody:

Primarily, the envisaged data repository must find its abode in an Object Storage medium, characterized by maximal interoperability with extant systems. This necessitates the exclusion of proprietary solutions such as Google Cloud Storage or Azure Storage Blob in favor of alternatives such as Hadoop Distributed File System (HDFS) [15] or any Object Storage infrastructure compliant with AWS S3 protocols. While HDFS presents itself as a viable candidate, the logistical intricacies entailed in deploying and administering a comprehensive HDFS cluster militate against prioritizing its adoption. Fortuitously, our discerning team has identified a panacea in the form of MinIO, an open-source Object Storage solution that not only aligns with the requisites of our project but also boasts compatibility with prevailing AWS S3 APIs.

The conundrum surrounding the optimal data format to be stored within the Object Storage reservoir has been a subject of fervent debate and contention amongst our learned colleagues. While proponents of denormalization advocate for its adoption to streamline data retrieval processes, empirical evidence underscores the efficacy of the RAW data format in compressing and storing multidimensional datasets, as attested by the diminutive file sizes characteristic of individual scan executions. Indeed, attempts to transmute this format into alternative structures, such as JSON, have yielded negligible dividends owing to the intrinsic complexity of the dataset. Moreover, with the ascendancy of multimodal Machine Learning paradigms, necessitating the simultaneous retrieval of diverse data fields to facilitate batch processing, the retention of all pertinent fields within a singular RAW file emerges as a prescient strategy poised to accommodate the evolutionary trajectory of our modeling endeavors.

In delineating a nomenclatural schema for the files ensconced within the Object Storage repository, precedence is accorded to a rudimentary ordering scheme premised on tenant identification and temporal demarcation. For the myriad utilization scenarios contemplated hereto-

fore, the concatenation of the tenant identifier with a timestamp proves to be a judicious nomenclatural convention. Furthermore, the hierarchical organization of files within the Object Storage framework ought to be predicated on a delineation between tenant-specific directories and temporal partitions.

Regarding the temporal component of the nomenclatural schema, adherence to the ISO-8601 standard, a globally ratified temporal convention, is deemed imperative. The employment of the ISO-8601 timestamp format, typified by the concatenation of year, month, day, hour, minute, and second components devoid of any superfluous delimiters, serves to obviate compatibility issues with extant filesystems. Notably, the conspicuous absence of special characters such as hyphens and colons serves to circumvent prohibitions imposed by certain filesystems, such as Windows, on the usage of such characters within filenames.

3.3.2 Data modeling for business processes

To provide a clear and logical understanding of this process, we will explore three critical models that collectively define the architecture of modern software systems: core model, identity model, and form model. Each model serves a distinct purpose and complements the others, ensuring a robust, secure, and user-friendly application. Let's delve into these models in the order that mirrors their integration and importance in the development process.

We begin with the core model, the backbone of our application. This model is where the fundamental business logic and core data structures are established. It defines the essential entities and their relationships that form the basis of the system, focusing on how data is organized and manipulated to support the application's primary services and functionalities. By starting with core model, we set the groundwork for all other aspects of the system, highlighting the primary operations and data flows that are crucial for the application's purpose.

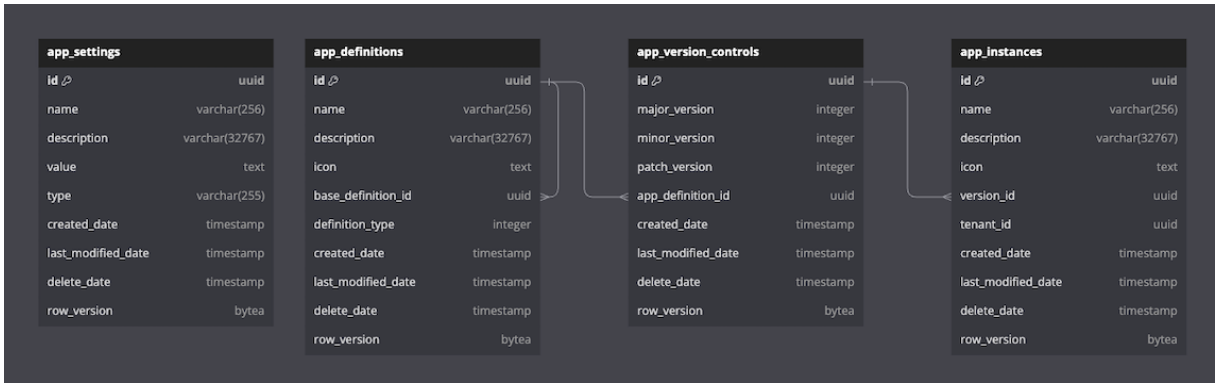


Figure 3.3: *The Core Model*

Once the core functionalities are established, we shift our focus to security with the identity model. This model is dedicated to managing user authentication and authorization. It ensures that access to the information and functionalities defined in core model is securely controlled. By integrating Identity model, we discuss how the system protects sensitive data and operations from unauthorized access, ensuring that users can interact with the core components in a secure environment. This model not only reinforces the security of the system but also defines user roles and permissions, which are essential for a multi-user application.

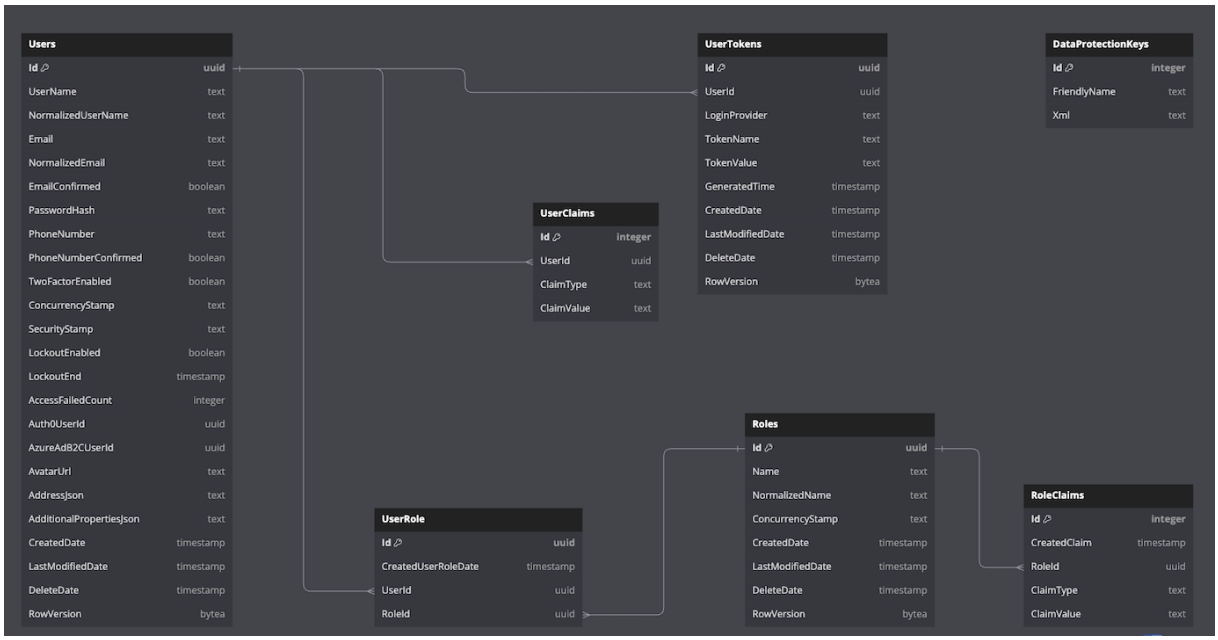


Figure 3.4: *The Identity Model*

Finally, we explore the form model, which ties the user interface and interaction directly to the underlying data structures of the core model and the security protocols of the Identity model.

This model addresses how users will experience and interact with the application through various forms and interfaces. It encompasses everything from data entry and management to ensuring that user interactions are intuitive and efficient. form model is critical for defining the presentation and accessibility of the system's features, making sure that the application is not only functional and secure but also user-friendly and responsive to the needs of its users.

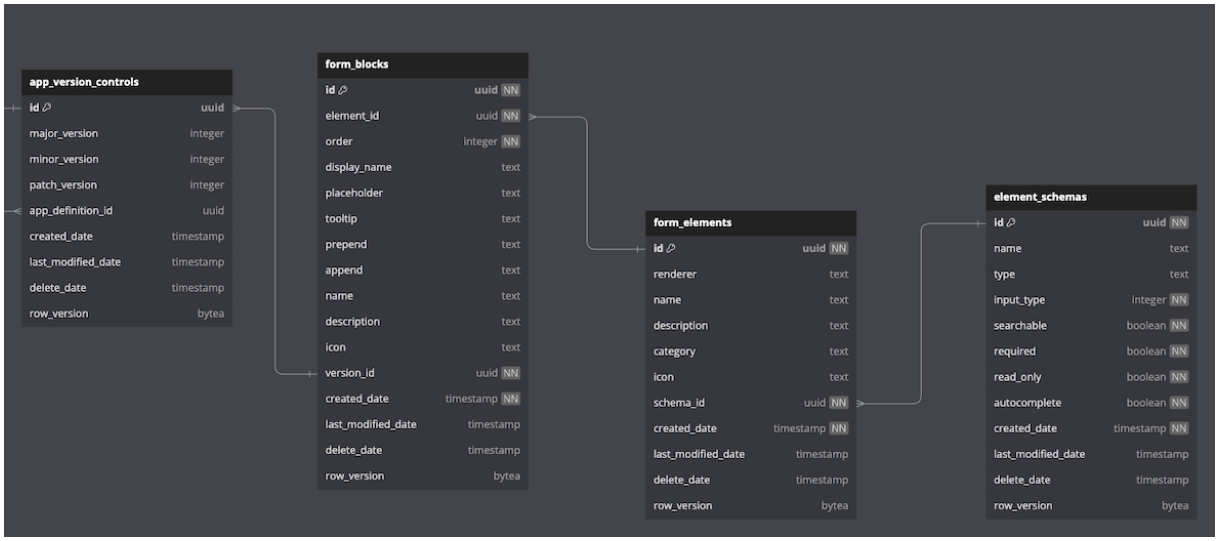


Figure 3.5: *The Form Model*

By examining these models in this order, we effectively build from the internal architecture to the external user interface, providing a comprehensive view of the system's design from the ground up. This approach helps in understanding how each component is interlinked and why each is vital to the overall success and operability of the application.

3.3.3 Data Layer of User Demands

The data layer of user demand in MeteorFlow encompasses various components and processes designed to handle different types of data submissions and interactions. This layer is divided into two main categories: Lay Users and Power Users, each with distinct data sources and processing mechanisms.

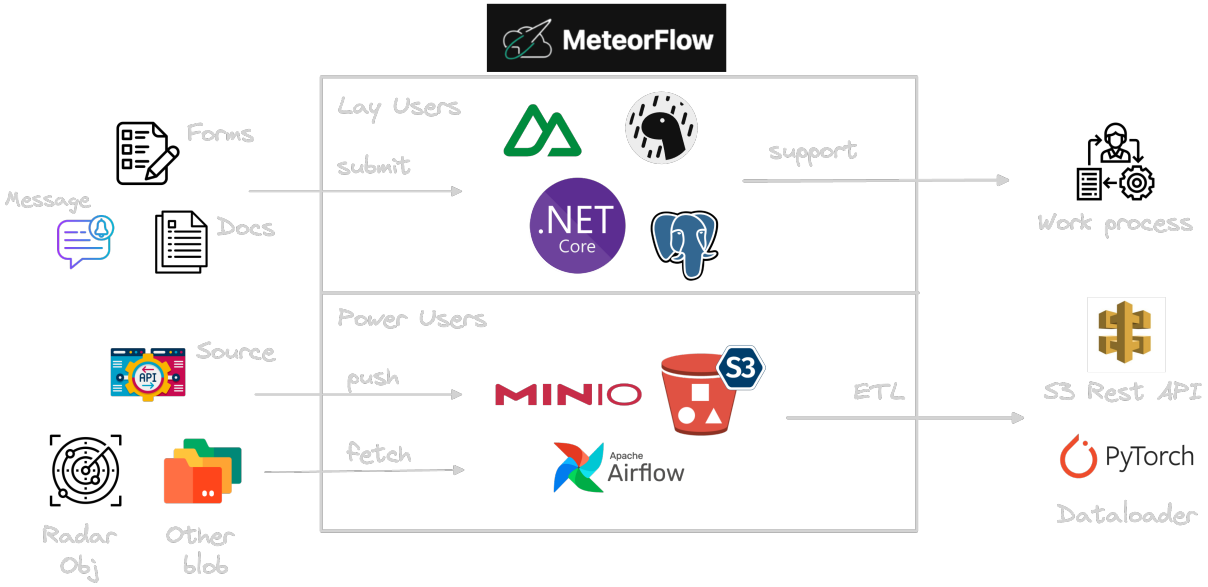


Figure 3.6: *The Form Model*

Lay Users

Lay users interact with the system primarily through forms, messages, and documents. These interactions are facilitated and supported by a combination of frontend and backend technologies.

Forms submission for lay users is managed through a dynamic and interactive user interface built with Nuxt.js. The submitted forms are then processed on the backend using .NET Core and PostgreSQL, ensuring efficient data collection and storage. This robust setup supports the overall work process by enabling seamless data management.

For message handling, the frontend integrates with various messaging platforms to facilitate user notifications and communication. These messages are managed on the backend using .NET Core, ensuring effective communication within the work process. Document management is another critical aspect, with Nuxt.js providing interfaces for handling documents, while .NET Core and PostgreSQL manage storage and retrieval. This setup ensures that necessary documents are easily accessible, supporting the overall work process.

Power Users

Power users require more advanced data handling capabilities, particularly for tasks involving data ingestion, storage, and processing. This is managed through the ETL (Extract, Trans-

form, Load) process.

Source data handling for power users involves fetching data from various sources via APIs. The data is then stored in MinIO and S3, accommodating large datasets and blob storage needs. Apache Airflow orchestrates the ETL process, ensuring data is correctly extracted, transformed, and loaded into the system. Specialized data types, such as radar objects and various other blobs, are also managed and stored efficiently.

Advanced processing and dataloading tasks are handled using PyTorch, which provides powerful capabilities for data analysis and machine learning. This setup is complemented by the S3 REST API, facilitating seamless interaction with S3 storage for efficient data retrieval and management.

Achievement

The data layer of user demand in MeteorFlow is designed to be comprehensive and versatile, accommodating the needs of different user types through a combination of modern technologies and streamlined processes. This structure ensures that data is handled securely, efficiently, and effectively, supporting the overall goals of the MeteorFlow system.

3.4 System Architecture

In today's rapidly evolving technological landscape, building robust, scalable, and maintainable systems is paramount. A well-designed system architecture not only addresses these needs but also provides a foundation for future growth and adaptability. This section delves into the architecture that underpins our system, exploring the principles of microservices, the application of Clean Architecture, and the integration of various infrastructure components to create a cohesive and efficient solution.

3.4.1 Microservices

In essence, a microservices architecture decomposes a large software application into a collection of smaller, self-contained services. Each service plays a specific role and owns a well-defined business capability. These services are loosely coupled, meaning they interact through well-designed APIs and operate independently.

Microservices architectures offer a compelling approach to building software. By decomposing large applications into smaller, independent services, they unlock agility, maintainability, and fault isolation. Each service owns a specific business function and interacts with others through well-defined APIs. This allows for faster development cycles, easier maintenance of individual services, and the freedom to choose the best technology for each job.

However, this power comes with a price. Managing a distributed system with numerous services inherently increases complexity compared to a monolithic application. Communication between services adds overhead to the system's performance. Perhaps the most significant challenge lies in ensuring data consistency across these independent services, requiring meticulous design and implementation. Carefully considering these trade-offs is crucial before embarking on a microservices journey.

While microservices boast impressive agility and maintainability, adopting this architecture isn't without its complexities. Managing a distributed system with numerous services inherently increases complexity compared to a monolithic application. Furthermore, communication between these services adds overhead to the system's performance. Perhaps the most significant challenge lies in ensuring data consistency across these independent services, requiring meticulous design and implementation.

3.4.2 Clean Architecture

Microservices and Clean Architecture can complement each other to create a modular, maintainable, and scalable system. Microservices architecture promotes the development of small, independent services that can be deployed, scaled, and updated independently. Each microservice encapsulates a specific business capability or domain, reducing the overall complexity of the system.

Clean Architecture, on the other hand, advocates for a layered approach to software design, promoting separation of concerns and decoupling of components. By adhering to the principles of Clean Architecture within each microservice, developers can achieve a high degree of modularity, testability, and maintainability.

Clean Architecture separates the system into distinct layers, each with specific responsibilities. The Shared Kernel acts as the core, housing reusable components that benefit all layers. The Domain layer sits at the heart of the application, defining core entities and their behaviors.

A common base class promotes code maintainability within this layer. The Application layer orchestrates request routing and establishes project-wide contracts for implementation in other layers. Clean Architecture emphasizes the testability of each layer through unit tests. Frameworks like xUnit simplify unit test creation.

The Infrastructure layer provides supporting services for the application. Cross-cutting concerns like logging reside here. User registration, authentication, and authorization functionalities are handled by the Identity layer. Persistence takes care of data access using patterns like repositories and Unit of Work. The Web Framework layer manages configurations for the web application, while the Web API layer delivers functionality to the user. Finally, plugins bridge the gap between monolithic and microservices architectures by promoting modularity.

3.4.3 Recommended System Architecture

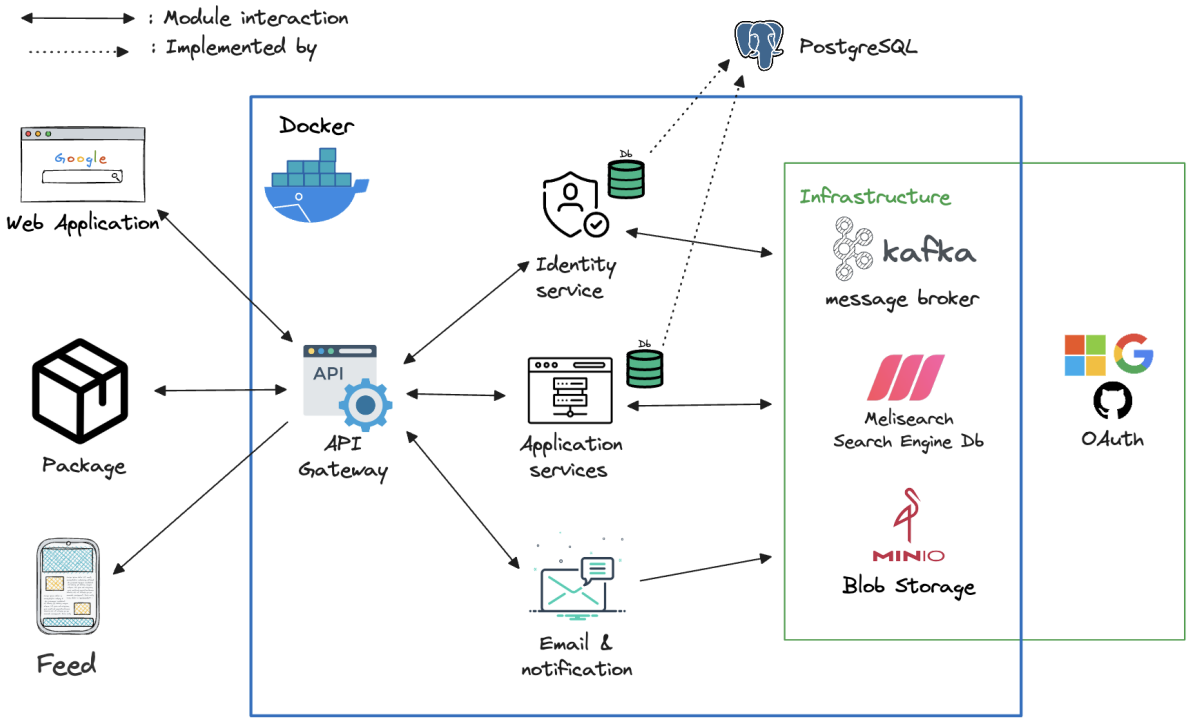


Figure 3.7: *System Architecture*

Based on the requirements set by stakeholders and after studying the existing system, the team proposes the design and implementation of a Weather Data Platform. Figure 3.7 illustrates

the implementation of the system.

Central to this architecture is the use of Docker, which encapsulates the application's microservices in containers. This encapsulation ensures that each service can be deployed and scaled independently across different environments without compatibility issues. Docker not only facilitates easy deployment but also enhances the manageability and scalability of the application services, making the infrastructure robust and flexible.

Within this architecture, a critical component is the Identity Service. This service is responsible for managing user authentication and authorization processes. It likely operates with a dedicated database that stores user credentials and session information, ensuring secure and efficient user access control. By centralizing identity management, the system can enhance security and simplify the integration of different services that require user identification and access control.

The Application Services form the backbone of the system's business logic. These services handle the core functionalities of the application, interacting with the Identity Service for authentication purposes and accessing dedicated databases to retrieve or store data. This separation of concerns allows for better maintenance and scalability of the application logic, enabling each service to be optimized and scaled according to specific demands without affecting the overall system performance.

An Email & Notification Service is integral to the architecture, managing communications with users. This service automates the sending of emails and other notifications, which are crucial for user engagement and timely communication. By having a dedicated service for this function, the system ensures that notifications are handled efficiently, maintaining high performance even under heavy loads of communication tasks.

The architecture is further supported by robust infrastructure components. PostgreSQL serves as the primary database management system, offering reliable and efficient management of structured data with its powerful SQL capabilities. Kafka, used as a message broker, ensures seamless and reliable data flow between different parts of the application, which is essential for maintaining data consistency and decoupling services. Meilisearch enhances the application's functionality by providing a fast and responsive search engine database, which significantly improves the user experience through quick search results. MinIO offers a high-performance solution for blob storage, managing unstructured data such as media files, backups, and logs, which

supports the application's scalability and data management needs. Lastly, OAuth is employed to facilitate secure delegated access, allowing the application to authenticate users by integrating with external OAuth providers like Google, thus broadening the scope of user accessibility and security.

Each component of this architecture plays a vital role in ensuring the application is scalable, modular, and resilient, employing a combination of state-of-the-art tools and technologies to meet a broad range of operational requirements efficiently.

Chapter 4

Implementation

4.1 Data Processing

The implementation phase of the team will be an extension to the current system, which is outlined in Figure 3.7.

In step 3, the team will set up a simple SFTP server. SFTP is a straightforward and widely used protocol, supported by numerous libraries and tools for communication based on this protocol. Additionally, compared to FTP, the mentioned protocol ensures security during data transfer. Depending on the permissions, the team may assist the observation station in constructing scripts to automatically forward processed files or allow manual file submission.

Once files are uploaded to the SFTP server, the team utilizes Airflow to orchestrate all existing ETL workflows in the overall system. Currently, the team pauses with a single DAG to process data from the Nha Be observation station. Airflow monitors newly added files on our SFTP server and initiates the ETL process. The choice of Apache Spark is based on the volume and complexity of the data. If the data size per new SIGMET file is manageable with Python alone, without the need for Spark, it will not be employed in this step.

Meteorological data, upon reaching the team's infrastructure, will be bifurcated into two main streams: Metadata, such as creation date, size, timestamp, etc., will be stored in a traditional Relational Database Management System (RDBMS). Specifically, PostgreSQL is chosen due to its popularity and the team's familiarity. Storing metadata here facilitates rapid query responses without direct access to the raw data. Common queries may include:

- What timestamps are being recorded? (e.g., from 21/11/2023 to 17/12/2023)

- At timestamp x, what are the geographical coordinates of the radar?
- What fields of data are currently stored?

Additionally, the database acts as an index, quickly identifying the storage location of raw data.

For specific hydrometeorological data, the team finds it inefficient to store them directly in conventional DBMS. Simultaneously, storing data in files still maintains a reasonable overall size. Therefore, the team decides to separate the raw data and store it directly in files. This approach, combined with the previously mentioned indexes, accelerates the retrieval process.

To facilitate data queries for models, machine learning, AI, etc., the team will develop a simple backend server using Python's FastAPI at step 5. In step 6, the backend receives query data in REST API format, queries the metadata DB and data files, and returns the achieved results. During different training instances, Machine Learning entities can connect to this server to retrieve data.

It's worth mentioning that the entire system will be developed and operated in a containerized manner and will be deployed on the Kubernetes platform. This reflects the system's ability to maintain high availability and ease of solution maintenance. In this illustration, the team will deploy it on a cluster of Raspberry Pi-embedded computers.

Lastly, in step 7, the team proposes an additional consideration. If suitable, the team may build a DataLoader to swiftly serve other model-making groups. Considering the popularity of Pytorch in AI, the team will initially approach this platform.

4.2 System

4.2.1 Programming languages and Libraries

C# was selected for this project because of its widespread popularity in enterprise applications and its capability to operate across various operating systems following the release of .NET Core. This feature significantly enhances deployment flexibility, allowing the application to be utilized in diverse environments.

Additionally, we have incorporated ASP.NET Core specifically version 8 into our technology stack. By leveraging ASP.NET Core, we benefit from a robust, well-supported framework

that facilitates the development of scalable and secure web applications. It seamlessly integrates with C#, enabling us to utilize a consistent programming environment while also exploiting features such as dependency injection, a vast ecosystem of middleware, and a strong configuration system that is suited to modern web applications.

ASP.NET Core 8 brings forward improvements in areas such as minimized startup times, reduced memory footprint, and enhanced security features, making it an ideal choice for developing scalable and secure web applications. The choice also underscores our commitment to developing applications that are both efficient and future-proof, ensuring that they perform optimally on both Windows and non-Windows platforms. This alignment with .NET Core's cross-platform capabilities ensures that our project remains versatile and adaptable to the evolving technological landscape.

Optionally, we have integrated the Nuxt framework into our technology stack. Nuxt is a progressive Vue framework that is used for building more robust and versatile web applications. It simplifies the development process by handling various aspects of the web infrastructure, such as server-side rendering, static site generation, and automatic code splitting. This inclusion enriches our application's interactivity and user experience, providing a seamless and dynamic interface for users.

4.2.2 Command Query Responsibility Segregation

The Command Query Responsibility Segregation (CQRS) is an architectural pattern that distinctively separates the tasks of reading data (queries) and writing data (commands) within a software application. This separation splits responsibilities into two main components:

- **Command Side**: This component manages operations that modify the system's state. It handles incoming commands from clients or external systems, conducts validations, and updates the data store accordingly. This side is essential for maintaining the integrity and accuracy of data modifications within the application.
- **Query Side**: Dedicated to data retrieval, this component processes all read requests. It fetches data from the appropriate sources, ensuring that the information provided is accurate and reflects the current state of the data store.

Key advantages of employing the CQRS pattern are:

- **Scalability**: CQRS allows for the independent scaling of the read and write components based on their respective workloads, which can significantly enhance system performance.
- **Flexibility**: With the separation of concerns, different storage and optimization strategies can be applied to the reading and writing processes. This flexibility enables the use of the most appropriate tools for each function, optimizing efficiency.
- **Event-Driven Architecture Compatibility**: The use of CQRS often complements event-driven architectures, where changes in the system's state are captured and managed as events. This compatibility ensures that the architecture is dynamic and responsive to changes in business requirements.

4.2.3 Project Structure

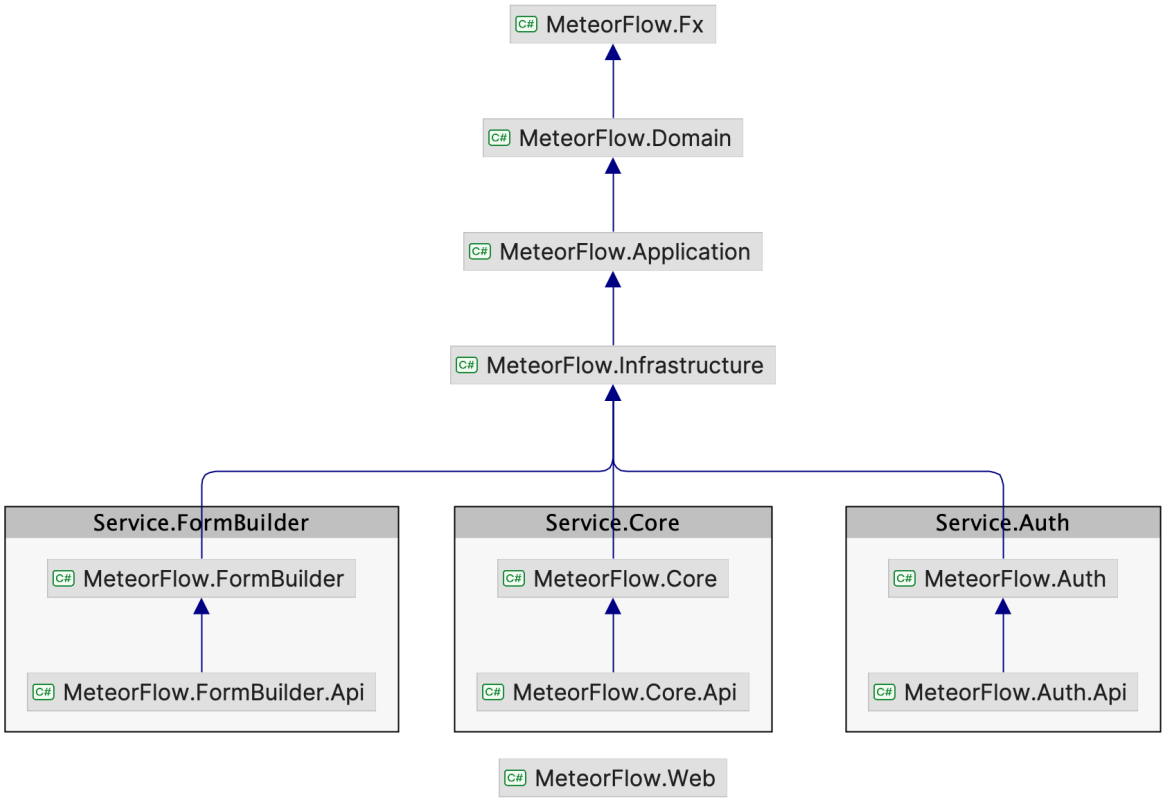


Figure 4.1: *Project Structure*

MeteorFlow is depicted as a modular framework comprising several interdependent components. Each library or module serves a distinct function, collectively supporting a robust and

scalable application infrastructure. This section will delineate the roles and relationships of these components.

MeteorFlow.Fx enriches the application by introducing additional functionalities or user interface enhancements. These features, while not central to the primary business logic, significantly augment the application's overall capabilities, providing enriched user experiences and functional extensions.

MeteorFlow.Domain is tailored to articulate the business domain, encapsulating entities and rules essential to the application's domain logic. Its reliance on the Core module underscores the latter's foundational role within the architecture, affirming its influence over the domain-specific functionalities.

At the heart of the architecture, MeteorFlow.Core embodies the fundamental business logic and operations critical to the application. Its pivotal role is underscored by its influence on all peripheral modules, which depend on the Core for their foundational functionalities.

MeteorFlow.Infrastructure functions as the backbone for data access and manages cross-cutting concerns such as logging and caching. This module is crucial for the operational management of the application, interfacing seamlessly with the Core module to apply essential functionalities across various infrastructural tasks.

Built upon the MeteorFlow.Infrastructure, the Auth module specializes in security and authentication processes, managing user verification and credential handling. The module provides APIs to enable external access to their functionalities and seamlessly integrate with other systems and applications. Other modules currently or in the near future follow the same structure above to provide additional functionalities.

MeteorFlow.Web, primarily tasked with managing the application's API gateway, interacts with all other modules within MeteorFlow to ensure robust web operations and effective resource management. Its role is crucial in maintaining the integrity and performance of web-based services, underpinning the system's interaction with users and external systems.

4.3 API Specification

4.3.1 Core API

GET /api/core/definition

Tags	Definition
Description	Retrieves a list of all core definitions.
Responses	200: An array of core definitions returned successfully. 401: Unauthorized access.

POST /api/core/definition

Tags	Definition
Description	Creates a new core definition.
Request Body	AppDefinitions schema.
Responses	201: Core definition created successfully. 400: Bad request if the data is invalid. 401: Unauthorized access.

GET /api/core/definition/{id}

Tags	Definition
Description	Retrieves a specific core definition by ID.
Parameters	id: UUID of the core definition.
Responses	200: Core definition returned successfully. 404: Core definition not found. 401: Unauthorized access.

DELETE /api/core/definition/{id}

Tags	Definition

Description	Deletes a specific core definition by ID.
Parameters	id: UUID of the core definition.
Responses	200: Core definition deleted successfully. 404: Core definition not found. 401: Unauthorized access.

POST /api/core/instance

Tags	Instance
Description	Creates a new instance.
Request Body	AppInstances schema.
Responses	201: Instance created successfully. 400: Bad request if the data is invalid. 401: Unauthorized access.

GET /api/core/instance/{id}

Tags	Instance
Description	Retrieves a specific instance by ID.
Parameters	id: UUID of the instance.
Responses	200: Instance returned successfully. 404: Instance not found. 401: Unauthorized access.

DELETE /api/core/instance/{id}

Tags	Instance
Description	Deletes a specific instance by ID.
Parameters	id: UUID of the instance.
Responses	200: Instance deleted successfully. 404: Instance not found.

	401: Unauthorized access.

GET /api/core/setting

Tags	Setting
Description	Retrieves a list of all settings.
Responses	200: An array of settings returned successfully. 401: Unauthorized access.

POST /api/core/setting

Tags	Setting
Description	Creates a new setting.
Request Body	AppSettings schema.
Responses	201: Setting created successfully. 400: Bad request if the data is invalid. 401: Unauthorized access.

GET /api/core/setting/{id}

Tags	Setting
Description	Retrieves a specific setting by ID.
Parameters	id: UUID of the setting.
Responses	200: Setting returned successfully. 404: Setting not found. 401: Unauthorized access.

DELETE /api/core/setting/{id}

Tags	Setting
Description	Deletes a specific setting by ID.
Parameters	id: UUID of the setting.

Responses	200: Setting deleted successfully. 404: Setting not found. 401: Unauthorized access.

GET /configuration

Tags	FileConfiguration
Description	Retrieves the current configuration.
Responses	200: Configuration returned successfully. 401: Unauthorized access.

POST /configuration

Tags	FileConfiguration
Description	Updates the current configuration.
Request Body	FileConfiguration schema.
Responses	200: Configuration updated successfully. 400: Bad request if the data is invalid. 401: Unauthorized access.

4.3.2 Indentity API

GET /api/auth

Tags	Auth
Description	Retrieves authentication status based on the returnUrl parameter.
Parameters	returnUrl (string): The URL to return to after authentication.
Responses	200: Authentication status returned successfully.

POST /api/auth/login

Tags	Auth
Description	Logs in a user with provided credentials.
Request Body	LoginInfo schema: Includes username and password.
Responses	200: Login successful.

POST /api/users/{id}/passwordresetemail

Tags	Users
Description	Sends a password reset email to the user specified by ID.
Parameters	id (string, uuid): Unique identifier of the user.
Responses	200: Password reset email sent successfully.

PUT /api/users/{id}/password

Tags	Users
Description	Updates the password for the user specified by ID.
Parameters	id (string, uuid): Unique identifier of the user.
Request Body	PasswordSetter schema: Includes new password and confirmation.
Responses	200: Password updated successfully. 404: User not found.

POST /api/users/{id}/emailaddressconfirmation

Tags	Users
Description	Sends an email confirmation to the user specified by ID.
Parameters	id (string, uuid): Unique identifier of the user.
Responses	200: Email confirmation sent successfully.

4.4 Datastore

4.4.1 The purpose of storing data directly as raw file

In the comprehensive compilation acquired during the recent field excursion, an array of invaluable data was meticulously gathered, meticulously chronicled, and thoughtfully analyzed. Specifically, our focus gravitated towards the operational activities conducted within the radar center situated in the Nha Be district. Operating at the pinnacle of technological advancement, the radar center diligently executes periodic scans of the atmospheric conditions, facilitating the meticulous documentation and exportation of a SIGMET file, which serves as a repository encapsulating a myriad of meteorological phenomena and atmospheric dynamics.

The SIGMET file, serving as the quintessential embodiment of scientific rigor and methodological precision, meticulously records a plethora of critical parameters and meteorological variables. Among the salient features meticulously delineated within this archival masterpiece are the discerning metrics of reflectivity and wind radial velocity. Reflectivity, a fundamental metric in radar meteorology, is an indispensable parameter delineating the intensity of electromagnetic waves returned to the radar antenna from various atmospheric particles, thereby furnishing crucial insights into the spatial distribution and intensity of precipitation within the monitored region. Furthermore, the wind radial velocity, an elemental component in meteorological analysis, delineates the rate and direction of atmospheric motion along the radial axis relative to the radar antenna. These metrics serve as a vital tool in elucidating the intricate dynamics of atmospheric circulation, enabling meteorologists to decipher prevailing wind patterns, identify regions of convective activity, and forecast the trajectory of severe weather phenomena with enhanced accuracy and precision.

Yet, amidst the effervescent tapestry of meteorological data acquisition and analysis, a conundrum of paramount importance arises: the optimization of data storage and retrieval mechanisms. It is within this crucible of inquiry that the proposition to store meteorological results directly within the file repository assumes a mantle of significance, heralding a paradigm shift in data management practices.

Indeed, the inherent multidimensionality of meteorological data, encapsulated within its tripartite tensor structure, underscores the imperative for a streamlined approach to data storage and retrieval. By eschewing aggregation and denormalization techniques, we advocate for a

direct integration of results within the archival framework, thereby fortifying the foundations of data integrity and computational efficiency.

Moreover, the proposition to leverage the existing ETL (Extract, Transform, Load) system developed by Vaisala, as utilized by our esteemed National Center for Hydrometeorological Forecasting, emerges as a beacon of pragmatism and resourcefulness. In a landscape fraught with technological complexities and budgetary constraints, the utilization of pre-existing infrastructure represents a judicious allocation of resources, affording seamless integration and interoperability across disparate data management platforms.

In summation, the decision to store meteorological data directly within the file repository, without resorting to further aggregation techniques, emerges as a testament to both pragmatism and foresight. By embracing this approach, we not only enhance the accessibility and usability of meteorological datasets but also lay the groundwork for a new era of scientific inquiry and meteorological prognostication, fortified by the pillars of technological innovation and methodological rigor.

4.4.2 Extracting data from each of the radar center

We have authored a script, albeit not yet integrated into the operational system, designed to facilitate the redirection of the SIGMET (Significant Meteorological Information) file from the data center to our preconfigured Storage Server.

Presently, the script is implemented in Python to ensure platform-agnosticism. Accompanied by a concise setup script, it will be poised for execution on the radar center's machinery.

MinIO has been employed as an unstructured data repository owing to its compatibility with the S3 (Simple Storage Service) protocol. As the system matures in the distant future, transitioning to AWS S3 should be straightforward. The underlying logic and API for interfacing with the storage infrastructure remain largely invariant during such a migration.

```
file_path = pathlib.Path(args.filename)

s3_client = boto3.client( "s3", endpoint_url=os.getenv("S3_HOSTNAME"),
    aws_access_key_id=os.getenv("S3_ACCESS_KEY_ID"),
    aws_secret_access_key=os.getenv("S3_SECRET_ACCESS_KEY"), ) _ =
    s3_client.upload_file(file_path, os.getenv("S3_BUCKET"),
```

```
    generate_new_name(file_path.name))
```

Listing 4.1: Part of the script for uploading data to Storage

underfull hboxes

Currently, we follow a designated nomenclature for naming data files of the RADAR objects: `<year><month><day>T<hour><minute><second>`. This system enhances the speed of data retrieval for specific times of the day. Utilizing S3 Prefix filtering, we can efficiently select multiple data points within a time frame that may vary from a second to an entire year.

For instance, to query data for a specific day (e.g., 2024-03-15), utilizing the UNIX path wildcard, we can express our query as:

```
ls 20240315T*
```

Analogously, the same query can be executed using the `boto3` library in Python:

```
def query_with_wildcard(bucket_name): s3_client = boto3.client('s3')

    response = s3_client.list_objects_v2( Bucket='nha-be-radar',
        Prefix='20240315T' )

    return [obj['Key'] for obj in response['Contents']]
```

Listing 4.2: Querying data in 2024-03-15

Moreover, although our naming convention deviates somewhat from the ISO-8601 standard for timestamp representation, the inclusion of the letter **T** aids in swiftly identifying datetime components within the naming structure.

One limitation of this naming convention is its inability to support wildcard querying for a central time component. Consider the scenario necessitating data retrieval for a specific hour daily. Utilizing the UNIX wildcard, this can be depicted as:

```
ls 202403*T09*
```

The utilization of an internal wildcard here implies the absence of an efficient mechanism for querying such objects presently. The current recourse involves maximal prefixing (e.g., `Prefix=202403`), followed by subsequent filtering in Python.

4.4.3 Data Explorer

Besides S3-compatible API for accessing the storage, our solution also provides some alternative ways for preview the data storage. This can be used for manual intervention (in case of failure), or simply for the radar center's operators to preview.

First, there is an online explorer that can be access from the web browser. As you can see from Figure 4.2, the web UI is very user-friendly and easy-to-use. With the use of IAM (shorts for Identity and Access Management) and **Policies**, the admin user of our entire system can restrict what a tenant (or a radar center) can see and what should be hidden.

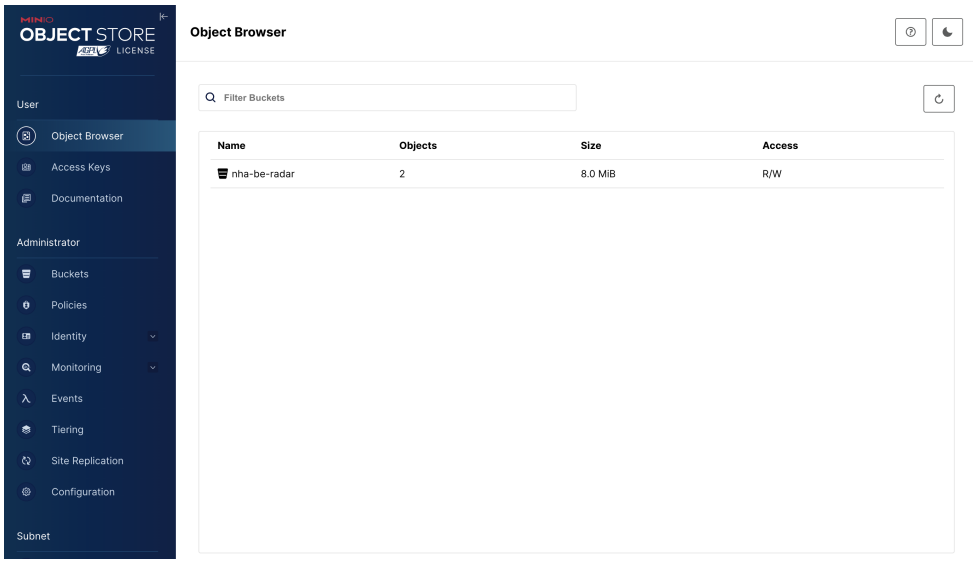


Figure 4.2: *MinIO web interface provides a clear and intuitive way to explore the storage*

Currently, as of the time of writing, our system has been self-hosted on a small Raspberry Pi server. The explorer has been set up with necessary load balancing, DNS and SSL/TLS registering correctly. Anyone that has been given proper permissions can visit `https://explorer.meteor-flow.com/` to see the server running. Figure 4.3 shows some of the files that our team has prepared beforehand.

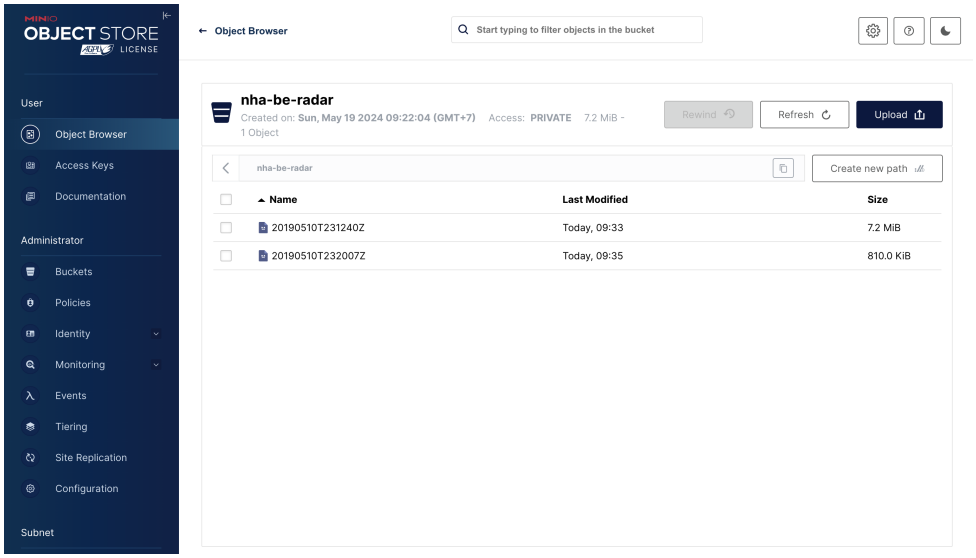


Figure 4.3: *Some radar files are currently being stored on the storage*

Not stopping at that, our system can also work with any other S3-compatible client. One notable example of this is Cyberduck. In the future, our team can research for more compatible protocols, such as FTP or SFTP.

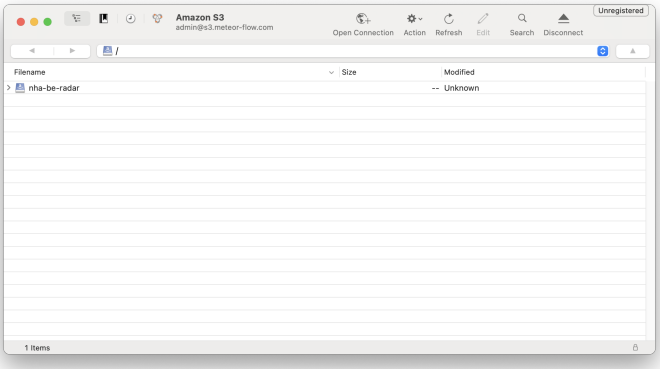


Figure 4.4: *Using Cyberduck to view data on the platform*

Chapter 5

Evaluation

5.1 Functionality

5.1.1 Object Storage and data preprocessing

The current status of our project involves the establishment of a rudimentary data lake infrastructure, meticulously crafted using robust and efficient Raspberry Pi servers. This infrastructure serves as the bedrock upon which we have meticulously implemented several essential features, akin to those prevalent in established data lake systems.

Foremost among these features is the implementation of a versatile gateway, meticulously designed to seamlessly ingest data from an array of incoming sources, ranging from radars to sensors. This gateway boasts compatibility with the Amazon Web Service S3-compatible API, rendering it universally accessible to clients and libraries engineered to interface with such protocols. Its compatibility with this widely adopted standard ensures that our gateway can seamlessly integrate with a plethora of external sources, thus enhancing the scalability and versatility of our infrastructure manifold.

Moreover, a robust platform for Data Orchestration has been deployed, with Apache Airflow serves as its cornerstone. This strategic implementation not only streamlines the current data management processes but also lays a solid foundation for the future scalability and expansion of our system. By orchestrating the scheduling, monitoring, and reporting of all data flows within the ecosystem, Airflow ensures the seamless integration of new tenants and providers, thereby safeguarding the integrity and efficiency of our data management operations.

In addition to these pivotal components, our team has diligently developed and published

PyTorch's native Data Loader to PyPI, a renowned package management platform within the Python developer community. This initiative represents a significant milestone in our endeavor to streamline data access and utilization within our ecosystem. By providing researchers with a user-friendly library that abstracts the complexities of low-level API interactions, we aim to significantly reduce the learning curve associated with navigating our data lake infrastructure. This strategic move not only enhances the accessibility of our system but also fosters a conducive environment for innovation and collaboration within the research community.

5.1.2 Form Builder and advanced features

The integration of the Form Builder as an add-on feature represents a strategic enhancement tailored to cater to the specific needs of our discerning customer base. Stemming from insights garnered during a comprehensive field trip to the Nha Be radar station, our team identified an opportunity to streamline and augment the operational workflow of operators tasked with responding to hazardous weather conditions. In response to this pressing need, we proposed and swiftly implemented the Form Builder feature, envisaging it as a pivotal tool in facilitating the seamless completion of warning forms amidst adverse weather events.

The preliminary implementation of the Building Form functionality marks an auspicious milestone in the development trajectory of this feature. While it remains in the nascent stages of development, we are optimistic about its potential to address the operational inefficiencies and pain points encountered by operators at the Nha Be radar station. By affording operators an intuitive and user-friendly interface for form completion, we envisage a tangible improvement in workflow efficiency and responsiveness, thereby bolstering the station's capacity to effectively manage hazardous weather conditions.

Furthermore, beyond its immediate utility as a workflow enhancement tool, the Form Builder feature holds promise as a source of valuable insights and data for our team. Through the aggregation and analysis of hand-labeled data generated during the form completion process, we anticipate gaining invaluable insights into the operational dynamics and decision-making processes underpinning hazard response scenarios. This rich repository of data not only informs iterative improvements to the Form Builder feature but also serves as a springboard for enhancing our understanding of user needs and preferences, thereby fostering a culture of continuous improvement and innovation within our organization.

Looking ahead, we remain steadfast in our commitment to refining and optimizing the Form Builder feature to meet the evolving needs and expectations of our stakeholders. Through ongoing collaboration and feedback from end-users, we endeavor to iteratively enhance the functionality and usability of this feature, ensuring its seamless integration into the operational workflows of operators at the Nha Be radar station. Moreover, we remain vigilant in our efforts to leverage the insights gleaned from this initiative to inform future product development endeavors, thereby solidifying our position as a trusted provider of innovative solutions tailored to the unique challenges of weather hazard response.

5.2 Performance Evaluation

Following the meticulous execution of rigorous load tests conducted on our self-hosted servers, we are pleased to present an insightful overview of the results obtained. Despite the deployment of our system on Raspberry Pi servers, it is imperative to underscore that this architectural choice has had minimal impact on the overarching efficiency and performance of our solution, as elucidated by the findings of our comprehensive assessment.

Upon subjecting our infrastructure to intensive load-testing scenarios, we observed a commendable level of resilience and stability exhibited by our Raspberry Pi-based servers. Contrary to preconceived notions regarding the limitations of such hardware configurations, our empirical data revealed that the performance metrics remained within acceptable thresholds, thereby affirming the viability of our chosen infrastructure for hosting data-intensive applications.

Furthermore, our meticulous analysis unearthed nuanced insights into the intricate dynamics governing the performance characteristics of our system. Leveraging sophisticated monitoring tools and methodologies, we meticulously scrutinized various performance parameters, ranging from CPU utilization and memory allocation to network throughput and latency. Through this meticulous examination, we gleaned a comprehensive understanding of the underlying factors influencing the efficiency of our infrastructure, thereby empowering us to optimize and fine-tune our system for enhanced performance and scalability.

5.2.1 Writing large volume of data

To rigorously evaluate the ingestion capabilities of our system, our team embarked on a meticulous testing regimen, leveraging the S3 interface for uploading Nha Be SIGMET files. Employing the Cyberduck client as our preferred tool for interfacing with the S3 protocol, we systematically endeavored to ascertain the system's ability to seamlessly ingest a substantial volume of data. The empirical findings gleaned from these rigorous tests furnish invaluable insights into the scalability and robustness of our infrastructure, thereby informing strategic decisions about its optimization and enhancement.

During our testing endeavors, we observed that the system demonstrated commendable resilience and efficiency, steadfastly accommodating the ingestion of up to 18 files in a single upload operation. While this threshold remained consistent across multiple test iterations, it is noteworthy that on certain occasions, our system exhibited heightened performance, facilitating the ingestion of approximately 25 files, with a cumulative size approximating 100 MB. This variability underscores the dynamic nature of our system's ingest capabilities, with performance metrics fluctuating in response to diverse environmental factors and operational parameters.

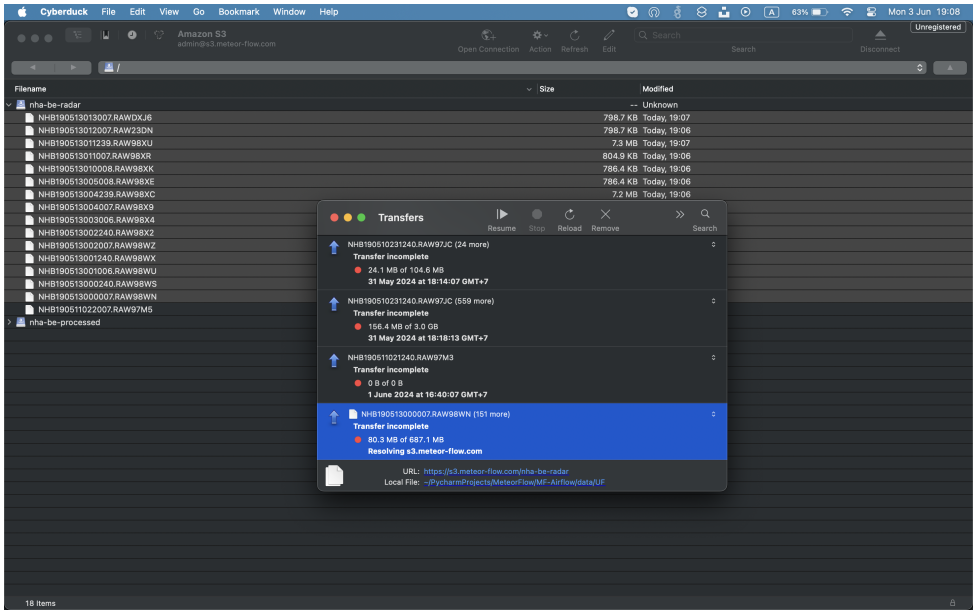


Figure 5.1: *Writing files in bulk*

Furthermore, it is imperative to contextualize these empirical findings within the broader framework of real-world data acquisition scenarios. Notably, the Nha Be radar generates meteorological data at a frequency of one scan every 10 minutes, thereby necessitating a robust

infrastructure capable of seamlessly handling the influx of data streams from disparate sources. Our empirical analysis revealed that our system is adept at managing up to 10 concurrent streams of incoming data, thus affirming its suitability for accommodating the diverse needs of tenants seeking to integrate with our ecosystem.

Moreover, it is essential to underscore the significance of these findings in the context of resource optimization and infrastructure scalability. Despite operating within the constraints of a low-end infrastructure, our system has demonstrated remarkable efficacy and resilience, underscoring the potential for leveraging cost-effective hardware solutions to achieve optimal performance outcomes. This strategic alignment of technological resources with operational objectives not only enhance the efficiency of our system but also augment its scalability and adaptability in response to evolving demands and requirements.

In conclusion, the comprehensive testing regimen undertaken by our team has yielded invaluable insights into the ingestion capabilities of our system, shedding light on its scalability, resilience, and performance characteristics. Through meticulous analysis and optimization efforts, we are poised to further enhance the efficiency and robustness of our infrastructure, thereby laying the groundwork for a future characterized by unparalleled innovation and excellence in data management and processing.

5.2.2 Processing data

In pursuit of enhancing the informativeness and query efficiency of our data, as well as mitigating loading bandwidth concerns, our team has judiciously opted to undertake preprocessing measures on the provided files. Central to this preprocessing strategy is the extraction of the "reflectivity" field, a pivotal attribute within the dataset, which is subsequently stored directly as a NumPy array. By meticulously orchestrating these preprocessing tasks within the purview of the Apache Airflow framework, we endeavor to streamline data manipulation processes while fostering a conducive environment for seamless data exploration and analysis.

The adoption of Airflow as our orchestration tool necessitated a concerted effort to optimize and customize the system to suit our specific requirements. Given the resource-intensive nature of Airflow, the initial deployment and configuration of a functional cluster posed a formidable challenge for our team. However, through a collaborative endeavor characterized by heavy optimization endeavors and meticulous customization efforts, we succeeded in overcoming these

hurdles and achieving a state of operational readiness. This transformative journey underscored the resilience and adaptability of our team, exemplifying our unwavering commitment to harnessing cutting-edge technologies for the betterment of our data management ecosystem.

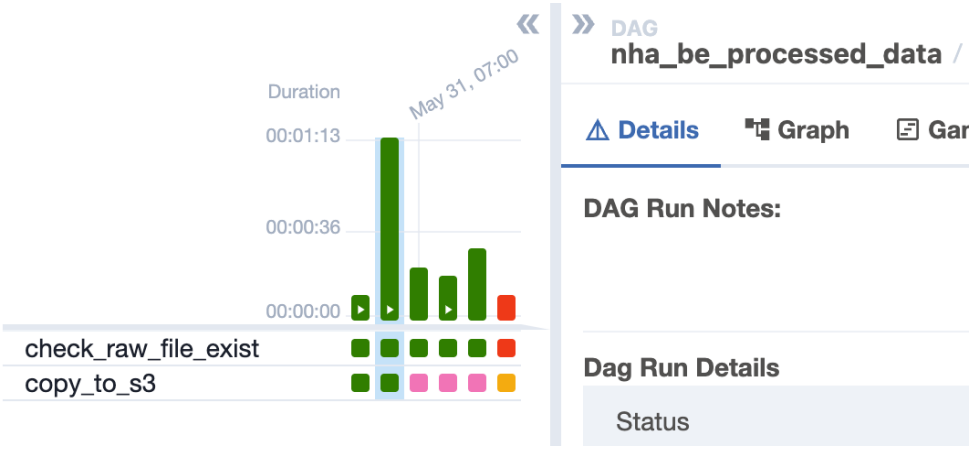


Figure 5.2: *It typically takes about 1.5 minutes to process one file from Nha Be*

It is imperative to delineate the tangible outcomes of our optimization endeavors within the broader context of operational efficiency and system performance. By fine-tuning the configuration parameters and streamlining the execution workflow, we have succeeded in substantially reducing the processing latency associated with data preprocessing tasks. Empirical evidence gleaned from extensive testing scenarios indicates that our optimized Airflow cluster can efficiently process a batch of five files, each approximately 30 MB in size, in less than two minutes. This remarkable feat not only underscores the efficacy of our optimization strategies but also augurs well for the scalability and resilience of our data processing pipeline.

Furthermore, it is essential to highlight the transformative impact of our preprocessing efforts on the overall efficacy of our data management ecosystem. By distilling the raw data into a concise and structured format, characterized by the direct storage of pertinent attributes as NumPy arrays, we have significantly enhanced the accessibility and usability of the dataset. Researchers and analysts can now leverage these preprocessed datasets to glean valuable insights and derive meaningful conclusions with unprecedented efficiency and accuracy, thereby catalyzing transformative advancements across diverse domains.

5.2.3 Concurrently reading data

Using the PyTorch API that we have implemented, we perform reading data in multiple processes. This innovative approach not only facilitates seamless access to our data storage but also serves as a poignant simulation of real-world scenarios wherein disparate model teams endeavor to interact with our comprehensive dataset.

Our empirical findings reveal that our servers exhibit commendable resilience and efficiency in handling the influx of querying threads accessing training data. Remarkably, our infrastructure demonstrates robust performance even under moderate load, with the capacity to seamlessly accommodate up to 5 querying threads without experiencing any detrimental impact on system stability or performance. This resilience is indicative of the robustness and efficacy of both the underlying MinIO storage system and our meticulously designed infrastructure architecture, affirming the soundness of our technological choices and engineering decisions.

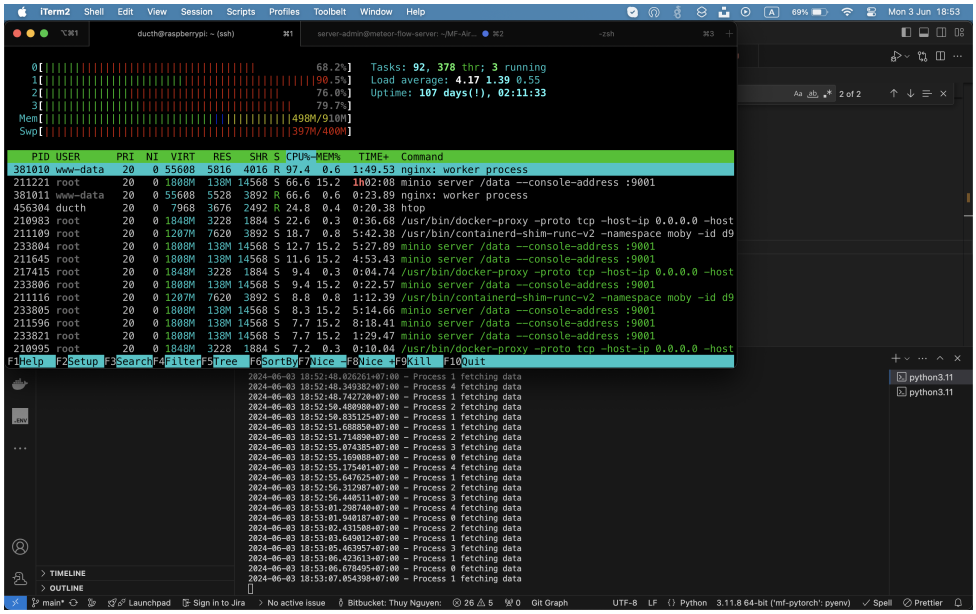


Figure 5.3: *Five concurrent thread reading data at the same time*

It is imperative to contextualize these empirical findings within the broader framework of model training workflows, wherein the interplay between data retrieval and model computation is pivotal. From a holistic perspective, the nominal increase in load time for each batch of data retrieval can be effectively mitigated by the inherent runtime required for model training between successive batches. Thus, while there may be a marginal increase in latency attributable to concurrent querying threads, this impact is mitigated by the inherent asynchronous nature of

model training workflows, thereby ensuring optimal utilization of computational resources and minimizing potential bottlenecks.

However, notwithstanding the commendable performance exhibited under current testing conditions, it is prudent to exercise caution and prudence in managing concurrent data access requests. Our empirical analysis suggests that while our infrastructure can robustly handle up to 5 concurrent readers accessing the data simultaneously, optimal system performance is ensured when the concurrency level is capped at 3 readers. By imposing this constraint, we can effectively safeguard the computational resources allocated for write operations, thereby enhancing system stability and ensuring uninterrupted data ingestion processes.

5.3 Further improvement

In the comprehensive review of our solution, it becomes apparent that while our system exhibits robust functionality and efficiency, there remain several areas where additional features or enhancements could be incorporated to further elevate its efficacy and usability. These missing features represent opportunities for refinement and optimization, underscoring the iterative nature of software development and the perpetual quest for excellence.

5.3.1 Supports more query types on PyTorch's API

Presently, within the framework of PyTorch's `MeteorFlowDataset`, data retrieval capabilities are confined to querying data within a specified range or interval of date-time parameters. For instance, users can query data pertaining to a specific year, such as 2019, or data corresponding to particular dates, such as the 10th and 11th of June. While this functionality suffices for basic data retrieval tasks, the burgeoning needs of model teams necessitate the expansion of query capabilities to encompass more sophisticated and nuanced requirements.

To address the evolving demands for enhanced data veracity, it is imperative to augment the existing query models with advanced functionalities that facilitate comprehensive data exploration and analysis. Among the proposed enhancements is the integration of support for querying data from combined areas encompassing two or more radar stations. This strategic augmentation not only facilitates the seamless integration of disparate data sources but also enhances the granularity and utility of the information returned.

The implementation of advanced query models, such as querying data from combined radar areas, represents a pivotal step towards enriching the capabilities of the `MeteorFlowDataset` and empowering model teams with unprecedented insights into meteorological phenomena. By enabling users to aggregate and analyze data from multiple radar stations simultaneously, we anticipate a quantum leap in the comprehensiveness and accuracy of the information gleaned from the dataset.

5.3.2 Increase cluster performance

The current deployment architecture of our system can be characterized as adequate, albeit with room for improvement. At present, the system is hosted on two Raspberry Pi 3B+ units, which effectively handle incoming user requests. However, despite their reliability in meeting current demands, our team has encountered challenges related to the intricacies of configuring and optimizing these ARM-based servers. The process of setting up the project entails meticulous installation of various configurations tailored specifically for these hardware platforms, often consuming significant time and effort, with execution times stretching over hours.

While the Raspberry Pi 3B+ units have served as dependable hosts for our services thus far, there is a palpable sense within our team of the limitations posed by their performance capabilities. As we strive for continuous improvement and scalability, it is imperative to explore alternatives that offer greater computational power and flexibility. Consequently, we envision conducting extensive research into the feasibility of transitioning to a more robust server solutions, such as Intel NUCs or basic PCs. By leveraging these higher-performance hardware platforms, we anticipate a significant enhancement in our system's capabilities, enabling us to seamlessly accommodate the anticipated growth in clusters and user base as we integrate with new tenants and systems.

The envisioned migration to more powerful server hardware represents a strategic investment in the future scalability and resilience of our infrastructure. Beyond the immediate benefits of enhanced computational capabilities and streamlined maintenance workflows, this transition holds the promise of positioning our system on a trajectory of sustained growth and innovation. With the agility and versatility afforded by higher-performance server solutions, we can confidently navigate the evolving landscape of data management and processing, ensuring seamless integration with emerging technologies and accommodating the dynamic needs of our expand-

ing user base.

Moreover, the adoption of alternative server hardware opens up avenues for exploring novel optimization strategies and performance enhancements. By capitalizing on the inherent strengths of Intel NUCs or basic PCs, we can unlock new possibilities for fine-tuning our system architecture and streamlining resource utilization. This strategic alignment of hardware capabilities with operational objectives underscores our commitment to continuous improvement and innovation, fostering a culture of excellence and resilience within our organization.

Chapter 6

Conclusion

The "MeteorFlow" Weather Data Platform (WDP) represents an ambitious endeavor to create a centralized, comprehensive solution for managing and utilizing weather data. Designed to cater to meteorologists, scholars, researchers, developers, businesses, freelancers, and NGOs, this platform aims to transcend traditional weather data offerings and deliver a multifaceted, immersive weather data experience.

The strategic growth of the MeteorFlow involves several pivotal enhancements and extensions to its initial design:

Non-linear Data: We plan to extend the data integration capabilities of WDP to encompass a richer variety of non-linear, detailed, and multi-source data. This will enable users to gain deeper insights into environmental conditions and patterns, significantly enhancing the scope and utility of the weather data provided.

Interactive User Interface: Our goal is to transform the user experience from mere data retrieval to an interactive engagement with weather forecasts. The interface will be designed to stimulate user curiosity and facilitate active interaction with the data, making the exploration of weather conditions more engaging and informative.

Geospatial Information Connection: Integrating geographical information systems, WDP will provide a localized, contextual view of weather forecasts. This feature is crucial for users needing to understand the specific impacts of weather on their immediate environment, enhancing both personal and professional decision-making.

Performance Optimization: Ensuring that the platform operates quickly and smoothly under various conditions is a priority. This will enhance user experience and ensure that the platform remains reliable and efficient, even under high demand.

Data Security: We commit to enhancing the security measures within WDP to protect the integrity and confidentiality of the data. This is essential for maintaining user trust and ensuring that sensitive information remains secure against potential cyber threats.

Deployment and Maintenance: Following deployment, WDP will adhere to a structured update cycle to consistently deliver precise and dependable weather information. Regular updates and maintenance will help adapt to evolving user needs and incorporate the latest technological advancements.

Through these development directions, the MeteorFlow aims not only to serve as a repository of weather data but also as a dynamic, evolving platform that responds to and anticipates the needs of its diverse user base. This will significantly contribute to the field of meteorology and related disciplines by providing a reliable, innovative tool for weather data analysis and application. In this study, we have successfully built a Proof-of-Concept with high applicability, aiming to streamline the steps in the conventional workflow. This is a crucial step towards optimizing and enhancing operational efficiency in both research and practical contexts.

Our goal was to create a flexible system with high adaptability, helping to simplify complex steps in the workflow. In doing so, we not only contribute to increasing efficiency but also alleviate the workload pressure on personnel, creating favorable conditions for creativity and focus on core tasks.

We didn't just stop at developing the system but also proposed flexible deployment strategies, emphasizing easy integration into the current working environment of those involved in information gathering and weather forecasting tasks.

References

[1] Airflow. What is airflow?, 2020. URL https://airflow.apache.org/docs/apache-airflow/stable/index.html. Last accessed by 16/12/2023.

[2] casey. Using range height indicator scan of radar, 2017. URL https://earthscience.stackexchange.com/questions/7222/using-range-height-indicator-scan-of-radar. Last accessed by 17/12/2023.

[3] Ramez A. Elmasri and Shamkant Navathe. *Fundamentals of Database Systems*. Addison Wesley, third edition, 1998.

[4] TRUNG TÂM DỰ BÁO KHÍ TƯỢNG THỦY VĂN QUỐC GIA, Jun 2203.

[5] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

[6] S. Hoyer and J. Hamman. xarray: N-D labeled arrays and datasets in Python. *Journal of Open Research Software*, 5(1), 2017. doi: 10.5334/jors.148. URL https://doi.org/10.5334/jors.148.

[7] Brenda Javornik, Hector Santiago III, and Jennifer C. DeHart. The lrose science gateway: One-stop shop for weather data, analysis, and expert advice. In *Practice and Experience*

in Advanced Research Computing, PEARC '21. ACM, July 2021. doi: 10.1145/3437359.3465595. URL http://dx.doi.org/10.1145/3437359.3465595.

[8] Kubernetes. Overview, 2019. URL https://kubernetes.io/docs/concepts/overview/. Last accessed by 17/12/2023.

[9] V. Lakshmanan et al. An improved echo top algorithm for radar meteorology. *Journal of Weather and Forecasting*, 28(2), 2013. URL https://journals.ametsoc.org/view/journals/wefo/28/2/waf-d-12-00084_1.xml. Available at: https://journals.ametsoc.org/view/journals/wefo/28/2/waf-d-12-00084_1.xml.

[10] G. Latisen and G. Vossen. *Models and Languages of Object-Oriented Databases*. Addison Wesley, 1998.

[11] Lrose. Radxconvert - lrose wiki, 2021. URL http://wiki.lrose.net/index.php/RadxConvert.

[12] Met Office. *Cartopy: a cartographic python library with a Matplotlib interface*. Exeter, Devon, 2010 - 2015. URL https://scitools.org.uk/cartopy.

[13] opengeospatial. Ogc standard netcdf classic and 64-bit offset. https://www.opengeospatial.org/standards/netcdf, 2011. Archived from the original on 2017-11-30. Retrieved 2017-12-05.

[14] 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/.

[15] Vineet Sajwan, Varnita Yadav, and M. Haider. The hadoop distributed file system: Architecture and internals. *International Journal of Combined Research & Development (IJCRD)*, pISSN:2321–2241, 05 2015.

[16] Michael Stonebraker and Uğur Çetintemel. 'one size fits all': An idea whose time has come and gone. In *Proceedings of the 21st International Conference on Data Engineering (ICDE)*, April 2005.

[17] Roland Stull. Weather Radars, 12 2022. [Last accessed by 17/12/2023].

[18] *RAW Product Format - IRIS Programming Guide - IRIS Radar*. Vaisala, 2024. URL `https://ftp.sigmet.vaisala.com/files/html_docs/IRIS-Programming-Guide-Webhelp/raw_product_format.html`.

[19] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.