EVENT-BASED DYNAMIC SENSOR CONFIGURATION FOR PERVASIVE ENVIRONMENTS

(YAYGIN ORTAMLAR IÇIN OLAY TABANLI DINAMIK SENSÖR YAPILANDIRILMASI)

by

İSMAİL OZAN ÇELİKEL, B.S.

Thesis

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

in

COMPUTER ENGINEERING

in the

GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

of

GALATASARAY UNIVERSITY

JUNE 2025

ACKNOWLEDGMENTS

I express my deepest gratitude to my supervisor Dr. Ozgun Pinarer for his support and guidance, this thesis wouldn't be possible without his wisdom and relentless support. Additionally, I want to thank all the faculty members of Galatasaray University for answering all my courious questions related and unrelated to my research.

I would also like to express my sincere gratitude to Professors Xavier Silvani and Christophe Paoli for allowing me to be a part of their project. Their guidance enabled me to apply my academic knowledge to a real-world system, an experience that significantly enriched my perspective. I am equally thankful to all the members of the SAPHiR Project at the University of Corsica for warmly welcoming me into their lab and making my time there both productive and memorable.

I am deeply grateful to my family and my friends for their unwavering support throughout this journey. Finally, I wish to thank my girlfriend for her endless patience, encouragement, and understanding, which have meant the world to me and kept me to push forward.

İsmail Ozan Çelikel
18 June 2025

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

In recent years, the use of Internet of Things (IoT) devices has grown rapidly, leading to the proliferation of IoT networks. As IoT adoption expands, the diversity of devices in use has also increased, resulting in significant heterogeneity across various dimensions, such as communication protocols and data transmission formats. While several methods have been proposed to address these challenges, the majority of existing research focuses on urban IoT applications. In contrast, rural environments typically exhibit greater device heterogeneity due to the wide range of environmental and operational conditions. Equipment selection in such settings is influenced not only by harsh weather and potentially hazardous environments, but also by factors such as energy efficiency and communication capabilities. These constraints limit the available options, making heterogeneity unavoidable. In this study, we propose a middleware solution designed for a real-time, heterogeneous IoT weather sensor network. The system delivers collected data to researchers, administrators, and end users. The middleware is carefully developed with attention to both semantic and syntactic differences among data sources and is integrated with an interface that facilitates access for various stakeholders.

Keywords : Data heterogeneity, rural IoT applications, IoT heterogeneity, weather sensors, meteorological data

RÉSUMÉ

Ces dernières années, la croissance rapide et l'adoption diverse des dispositifs IoT ont mené à des réseaux hétérogènes, notamment en termes de protocoles et de formats de données. Bien que la recherche IoT se concentre souvent sur les applications urbaines, les environnements ruraux présentent une hétérogénéité encore plus prononcée due à des conditions environnementales et des contraintes opérationnelles (efficacité énergétique, communication) qui limitent les choix d'équipement.

Dans cette étude, nous proposons une solution middleware pour un réseau de capteurs météorologiques IoT hétérogène et en temps réel. Ce système transmet les données aux chercheurs, administrateurs et utilisateurs finaux. Le middleware gère les différences sémantiques et syntaxiques des sources de données et intègre une interface pour un accès simplifié.

Mots Clés : Hétérogénéité des données, applications IoT rurales, hétérogénéité de l'IoT, capteurs météorologiques, données météorologiques

ÖZET

Son yıllarda, nesnelerin interneti (Internet of Things, IoT) cihazlarının kullanımı hızla artmış ve bu durum IoT ağlarının yaygınlaşmasına yol açmıştır. IoT'nin benimsenmesi arttıkça, kullanılan cihazların çeşitliliği de artmış ve bu durum, iletişim protokolleri ve veri iletim formatları gibi çeşitli açılardan önemli bir heterojenliğe neden olmuştur. Bu zorlukları ele almak için çeşitli yöntemler önerilmiş olsa da, mevcut araştırmaların büyük çoğunluğu kentsel IoT uygulamalarına odaklanmaktadır. Buna karşılık, kırsal alanlar genellikle daha geniş çevresel ve operasyonel koşullar nedeniyle daha yüksek düzeyde cihaz heterojenliği sergiler. Bu tür ortamlarda ekipman seçimi yalnızca zorlu hava koşulları ve potansiyel olarak tehlikeli çevresel etkenlerden değil, aynı zamanda enerji verimliliği ve iletişim protokollerinin efektifliği gibi faktörlerden de etkilenmektedir. Bu kısıtlamalar mevcut seçenekleri sınırlandırmakta ve heterojenliği kaçınılmaz hale getirmektedir. Bu çalışmada, gerçek zamanlı ve heterojen bir IoT meteoroloji sensörü ağı için tasarlanmış bir ara katman çözümü önerilmektedir. Sistem, toplanan verileri araştırmacılara, yöneticilere ve son kullanıcılara sunmaktadır. Geliştirilen ara katman, veri kaynakları arasındaki anlamsal ve sözdizimsel farklılıklar dikkate alınarak titizlikle tasarlanmış ve çeşitli paydaşların erişimini kolaylaştıran bir arayüz ile entegre edilmiştir.

Anahtar Kelimeler : Veri heterojenliği, kırsal IoT uygulamaları, IoT sistemlerde heterojenlik, hava durumu sensörleri, meteorolojik veriler

1 INTRODUCTION

In recent years, the idea of smart cities has gained a lot of attention from researchers, policymakers, and urban planners. They aim to use digital technology and data to solve the many problems that come with urbanization. However, as cities grow, it's becoming clear that rural areas also need these smart technologies. Rural communities face unique challenges related to the economy, environment, and infrastructure. Smart villages, which are smaller and more spread out than cities, are an important part of this larger smart city concept. These rural areas can greatly benefit from new data-driven methods that improve resource management, environmental sustainability, and the quality of life for residents.

Advances in technology, such as the Internet of Things (IoT) and Artificial Intelligence (AI), offer great potential to turn rural areas into sustainable smart villages. With more connected devices and sensors, we can collect real-time data on energy use, farming productivity, water usage, and environmental conditions in villages. AI can then analyze this data to provide useful insights, helping make better decisions and predict future needs. Given the rapid urbanization, climate change, and resource shortages, using IoT and AI in rural development is becoming increasingly important.

The frequency and intensity of extreme weather events are on the rise because of climate change (Intergovernmental Panel On Climate Change, 2007). This increased regularity puts rural environments at greater risk, as studies indicate that weather profoundly impacts these areas, affecting household income, food security, poverty, and overall welfare. Extreme weather events like droughts, floods, and temperature fluctuations reduce rural incomes, worsen poverty, and negatively impact nutrition and consumption. For instance, recent studies (Nguyen et al., 2020; Hussain et al., 2020) highlight the effects of floods and storms on agriculture and infrastructure in various rural examples. Similarly, research by (Mendelsohn et al., 2007; Gao and Mills, 2018; Coromaldi, 2020) illustrates diverse impacts from droughts and high temperatures. A study by (Deng et al., 2025) even suggests that weather shocks in certain areas influence loans provided by rural credit cooperatives, indirectly affecting the broader economy.

Given the significant and various impacts of extreme weather on rural areas, accurate weather prediction becomes critically important for local governments, decision-makers, and researchers. For local governments, precise forecasts enable the proactive implementation of disaster preparedness measures, resource allocation for emergency response, and the development of resilient infrastructure, safeguarding lives and livelihoods. Decision-makers in sectors such as agriculture, water management, and finance rely on timely and reliable weather information to make informed choices that mitigate economic losses and establish sustainable development initiatives. Researchers, in turn, leverage predictive models to better understand long-term climate trends, evaluate the effectiveness of adaptation strategies, and inform policy recommendations (Leu et al., 2014). The cornerstone of accurate weather prediction, however, lies in precise data collection. High-quality data from a robust network of sensors, satellite imagery, and other meteorological instruments is essential for feeding sophisticated forecasting models, reducing uncertainties, and providing the reliable insights necessary for effective planning and mitigation in the face of escalating climate challenges (Pandey et al., 2017; Xiao et al., 2012).

The main goal of this research is to explore how data-driven strategies can optimize the sustainability of smart villages. This includes using IoT and AI technologies to tackle the various challenges rural communities face. The research will design and set up an integrated IoT system to collect and analyze data from sensors, devices, and citizens in the smart village environment. Another goal is to develop AI-driven systems that use machine learning and predictive analytics to improve resource allocation, energy management, and risk reduction.

Additionally, this study will look at ways to manage weather monitoring in rural areas to help communities in these areas be more resilient, and be prepared for weather events. It will also use IoT sensors and AI models for real-time environmental monitoring and risk assessment to lessen the effects of natural disasters and climate change on smart villages. Moreover, the research aims to create interactive platforms and communication channels to engage the community, encourage citizen participation, and share knowledge within smart village projects, promoting collaboration and inclusivity.

Finally, the research will conduct thorough evaluations and impact assessments to measure the effectiveness and scalability of these data-driven solutions in real-world smart village deployments. This ensures that the benefits are real and significant for rural communities and stakeholders. By developing innovative IoT and AI-based solutions tailored to the specific needs of smart villages, this research aims to empower rural communities with data-driven decision-making tools and participatory governance, building local capacity and resilience.

This research also aims to advance interdisciplinary studies at the intersection of data science, engineering, and sustainable development, addressing critical challenges such as resource scarcity, environmental degradation, and socio-economic inequality. Moreover, it seeks to establish partnerships and collaborations with local stakeholders, government agencies, and international organizations to facilitate technology transfer and knowledge exchange for global impact. In summary, this research endeavors to unlock the transformative potential of data-driven optimization for sustainable smart villages, paving the way for inclusive, resilient, and thriving rural communities in the digital age.

As rural areas increasingly embrace digitization and smart technologies to enhance sustainability, there arises a pressing need for data-driven optimization approaches tailored to the unique challenges and opportunities of smart village environments. This study aims to investigate the synergistic integration of IoT and AI techniques to optimize resource management, environmental sustainability, and community well-being within smart villages and smart rural areas. By applying data-driven methodologies, the research seeks to develop innovative solutions that empower rural communities to make informed decisions and drive positive socio-economic and environmental outcomes. Through interdisciplinary collaboration and real-world deployments, this study endeavors to advance knowledge in the field of sustainable rural development while addressing practical challenges faced by smart villages worldwide.

The remainder of the study is structured as follows : Chapter 2 provides a comprehensive review of related work in this field, highlighting existing research contributions and gaps. Chapter 3 presents the approach and the methodology employed in this study, outlining the key definitions and assumptions. Chapter 4 introduces technical details of the study. Finally, Chapter 6 concludes the study by summarizing key findings, highlighting contributions, and outlining avenues for further exploration.

2 LITERATURE REVIEW

With the increasing use of technologies such as wireless networks, smart devices equipped with different sensors, radio frequency identification tags (RFID) and near field communication (NFC), the concept of the IoT, a new technology, has emerged in people's daily lives. There is interest in using wireless sensor technologies in various IoT scenarios. The idea of the internet of things offers a perspective that includes communication, data exchange, data collection and integration between the objects around us. Considering the tremendous growth of smart objects and applications, collecting and analyzing data therein emerges as the main challenge. It is critical that operations be energy efficient in battery-powered sensor nodes in IoT technologies. In networks formed by wireless sensor nodes (WSN), energy is largely consumed while data packets are received from neighboring nodes and transmitted to the central station. To reduce energy consumption in sensor networks, techniques can be applied to reduce the size of data packets by removing redundancies in transmitted or received data. Data aggregation is an important technique in getting rid of data redundancy and ensuring energy efficiency. It also extends the life of wireless sensor networks thanks to energy efficiency.

To deploy an optimal Wireless Sensor Network (WSN) that can respond to situations smartly and efficiently, several key concepts become crucial. These include data aggregation and communication in IoT. The sheer number of IoT devices within a network also leads to data heterogeneity, a significant variety in the data itself. This section will explore the importance of data aggregation within the IoT framework, delve into the implications of data heterogeneity and methods to address it, and finally, present recent studies on weather data systems.

2.1 Data aggregation

(Abbasian Dehkordi et al., 2020) introduces data aggregation techniques, which are an effective method in reducing energy consumption, which is the most important challenge of wireless sensor networks and are examined in depth in order to ensure the continuity and long life of the IoT system. Major techniques for data aggregation and integration in terrestrial, underground, underwater and body area wireless sensor networks are presented. In addition, the usage applications, advantages, disadvantages and difficulties of each technique are also explained.

The IoT allows any entity to send and receive data over a communication network. However, network lifetime is a critical challenge in IoT, affecting the longevity of IoT applications. To overcome this difficulty, data aggregation is used as an efficient method in networks where nodes are limited in terms of resources and energy. The main purpose of the data aggregation method is to effectively combine and aggregate data packets to improve energy consumption, network lifetime, traffic bottleneck and data accuracy. Additionally, removing redundancies in data and reducing the amount of data transfer also reduces energy consumption. Efficient collection of data depends on the network design and the size of the sensed data. Because the size of the data creates a burden on network communication.Although data aggregation plays a crucial role in IoT systems, there has not yet been a comprehensive and systematic study that organizes and analyzes the key techniques in this area. The study by (Pourghebleh and Navimipour, 2017) addresses this gap by reviewing current data aggregation approaches in IoT. In their work, data collection strategies are categorized into three main types : tree-based, cluster-based, and centralized aggregation. These classifications are based on how sensor nodes are organized in the network and how data is transmitted. In the tree-based approach, data gathered from the environment is sent from leaf nodes to intermediate nodes, where aggregation occurs before the data is forwarded to the sink node. In the cluster-based method, sensor nodes are grouped into clusters, each managed by a cluster head. Sensor data is first sent to the respective cluster head and then relayed to the nearest sink node. The centralized approach involves all sensor nodes sending their data directly to a central node via the shortest possible paths, where aggregation is performed and the result is sent as a single packet. The study also evaluates these methods using various criteria such as energy efficiency and latency. Furthermore, it discusses the key challenges and issues involved in implementing data aggregation in IoT environments.

(Alam, 2021) explore how cloud-based IoT applications enhance the functioning of smart cities by collecting and analyzing data from various sources. This data management helps improve urban infrastructure, including transportation, utilities, resource management, waste management, crime detection, and healthcare. Cloud computing, which offers IT capabilities as services, significantly boosts IoT efficiency and performance. By integrating IoT with cloud computing, smart cities can overcome challenges such as limited storage and power. Cloud-based platforms enable real-time data processing, helping cities optimize resource usage, reduce accidents, and minimize pollution. The study discusses the convergence of IoT and cloud computing, emphasizing their combined role in smart city development. Future research is suggested to uncover more applications and their impact on urban environments.

(Almalki et al., 2023) examine how the IoT can enhance smart cities while addressing environmental challenges like energy consumption, pollution, and e-waste. It advocates for "green IoT," aiming to create eco-friendly and sustainable urban environments. This comprehensive survey reviews techniques and strategies to make smart cities smarter, more sustainable, and eco-friendly. IoT in smart cities connects devices such as cameras, sensors, and mobile phones to improve various applications, including environmental monitoring, e-healthcare, and transportation. Efficient IoT techniques can reduce energy consumption and resource usage, contributing to greener cities. The paper discusses the need for advanced technologies like AI to further enhance smart city infrastructure and highlights future research directions for sustainable development. The study concludes by emphasizing the importance of green IoT in reducing pollution and energy use, making cities more sustainable and safer. It provides insights into eco-friendly IoT applications and future research opportunities to develop smarter, greener cities.

(Ashraf, 2021) underline the role of IoT devices in transforming traditional cities into smart cities. IoT devices collect vast amounts of data through wireless sensors, which is processed to improve city operations. The study highlights the use of IoT in managing big data and emphasizes the proactive role of IoT in creating smart cities. A significant focus is on smart traffic management systems, which use sensors and cameras to reduce traffic congestion by automatically adjusting traffic light durations based on real-time data. The Analytic Hierarchy Process (AHP) compares smart services like transport, energy, infrastructure, health, and agriculture in smart cities versus traditional ones. The results show that about 98% of people in smart cities are satisfied with their living conditions. The study highlights the need for technological preparedness as society progresses. It explains how IoT devices and sensors gather detailed information about city life, such as traffic and healthcare, to improve efficiency. The conclusion emphasizes the rapid development of IoT and its applications, including smart homes and agriculture. It stresses the importance of smart traffic systems in building smart cities and discusses tools for efficient big data management and analysis.

(Atitallah et al., 2020) discuss how the integration of IoT and Deep Learning (DL) technologies can address the challenges posed by rapid urbanization. IoT devices collect vast amounts of data that can be analyzed using DL techniques to create smart applications and services, enhancing the quality of life in cities. The review covers IoT's role in generating big data, the computing infrastructures used for data analytics (cloud, fog, and edge computing), and popular DL models applied to smart city contexts. The survey outlines various smart city applications, including healthcare, surveillance, agriculture, and transportation, and highlights the potential of DL to improve these sectors. It also discusses the challenges and opportunities in using DL for IoT big data analytics in smart cities. The article aims to provide a comprehensive overview of the current state of research in this field and identify areas for future exploration.

(Awan et al., 2020) conduct a comparative analysis of various Machine Learning (ML) and Deep Learning (DL) techniques to predict parking space availability using a dataset from Santander, Spain. The study evaluates well-known algorithms including Multilayer Perceptron, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, and Voting Classifier (Ensemble Learning). Results indicate that simpler algorithms like Decision Tree and KNN outperform more complex ones like Multilayer Perceptron in terms of prediction accuracy. The study highlights that Decision Tree emerged as the optimal solution, closely followed by Ensemble Learning. This research aims to enhance smart parking applications by integrating these predictive models and obtaining user feedback to refine and validate their effectiveness in real-world scenarios.

(Bauer et al., 2021) explore the development of Smart Cities, driven by increasing urbanization and the challenges it presents. It emphasizes the role of the IoT as a key technology enabling Smart City initiatives. The paper reviews various global Smart City projects that have integrated IoT to enhance urban living, highlighting the impact of these technologies on city services. It also discusses essential IoT technologies, their contributions, and the challenges that remain. The key issues include the need for interoperability, standardization, and data security. The study argues for the adoption of standard-based and open-source solutions to overcome the fragmented Smart City landscape and promote wider deployment and integration of Smart City services.

In IoT, there are different types of sensors and actuators that connect to the internet through different access networks, such as wireless sensor networks. Billions of objects can be found in IoT through these sensors and actuators. Thanks to the ability to connect objects to the Internet through sensors and actuators, IoT provides the opportunity to access physical objects in the environment anytime and anywhere. Wireless sensor networks (WSN) play an important and active role in the integration of information in the environment. In IoT applications, large amounts of data are sent from one point of the network to another. Sensor nodes are inherently energy limited as they operate on power such as batteries. Therefore, it is not effective and efficient in terms of energy consumption for sensor nodes to transmit all collected data directly to the sink node. Data generated from neighboring nodes is often correlated and highly redundant. By combining this related and redundant data with valid, high-quality information at intermediate nodes, the data collection method can be followed. This form of data collection has been shown to be beneficial in terms of power consumption, bandwidth usage, load balancing, network lifetime, and data accuracy. With proper data collection technique, energy consumption and network traffic volume can be reduced. In addition, it can increase network lifetime and data accuracy. (Rahman et al., 2016) compares several data collection techniques for some parameters considered important using the NS-2 simulation tool.And the obtained simulation results were analyzed and evaluated according to the selected performance measurements.

The IoT network consists of interconnected sensors, devices, actuators, and other connected nodes. These sensors and actuators connected in the IoT network are usually equipped with various types of transceivers, microcontroller devices and protocols necessary for control and communication of sensor data. Sensors in the network are connected to a central repository (server) to store measured data in real time and transmit the data there. Authorized persons access the data from this center. Therefore, without human intervention, IoT networks can operate safely and sustainably for a long time. This topology is called wireless sensor network (WSN). Heterogeneous WSN has become the cornerstone of IoT-based systems by connecting a wide variety of smart sensors. The rapid development of devices in WSN also brought energy consumption problems. Communication and data sharing in WSNs are the operations that cause the most energy consumption. A sensor node typically consists of 4 units. These ; the sensing/identification unit, which detects and collects data from the work area, the processing unit, where the detected data is manipulated, the communication unit, where the processed data is transmitted, and the power supply unit, which provides power to all components of the node. A sensor node can have 3 states based on its operating state ; active, sleeping and idle. According to these situations, the maximum energy consumption is in the active state. In the idle state, the node waits for a data packet to transmit data to another node. This causes energy consumption. Sleeping is the state with the least energy consumption since no action is taken and communication is turned off. Battery energy sources are used to power IoT systems. Therefore, efficient use and management of these resources is important. In this study, the role of WSN in IoT systems, its difficulties and data collection issues were investigated by conducting a literature review. (Gulati et al., 2022) present a research on the conservation of energy required for the operation of WSN and therefore the IoT system and various data collection techniques for this purpose are also presented.

With the influence of the Internet of Things, there has been a rapid growth in digital health applications, which has increased the efficiency of healthcare services and the quality of life of patients. For example ; vital values of the patient can be measured with IoT body devices. Wireless body sensor networks (WBSN) ; Wearable mobile devices and lightweight implantable medical devices help to diagnose the disease efficiently and produce better statistical results by measuring the patient's vital signs such as heart rate and blood pressure and presenting them to doctors. Since values such as body temperature, heart rate, and blood pressure are measured regularly and repeatedly with WBSN, the size of the data is very large. Therefore, these health data aggregation services play a critical role in terms of efficiency and verification. However, in wireless body sensor networks, a malicious attacker can obtain the patient's health status by eavesdropping on the communication network and accessing and analyzing real-time data on the IoT health device. Thus, it is important to protect the patient's data security and privacy. Encryption schemes are widely used in IoT body sensors to meet security and privacy requirements. A signature scheme is needed for data verification. Typically, data is digitally signed and then encrypted. However, an additional computational resource is needed for a scheme that performs both signature and encryption. Considering energy-constrained wearable devices, resource efficiency must be ensured when implementing these schemes. In this article, a new encryption accumulator based on a lightweight encryption scheme is proposed to enable secure collection of data from IoT body sensors in healthcare. This accumulator consists of pieces of data collected from sensors and is encrypted and signed before being sent to the data server. The data server cannot decrypt or modify the data, but it can compute the sum without decryption. In this way, the security feature of the data is ensured without the healthcare provider having to deal with data collection and aggregation. (Rezaeibagha et al., 2020) presents an IoT-based wearable monitoring system with cryptographic accumulator. In addition, the performance evaluation is made by comparing the computation and communication costs of this system with other IoT-based solutions.

Despite remarkable improvements in hardware and software technologies, there are many challenges in meeting the requirements of IoT applications in fields such as industry, agriculture, logistics, transportation, smart buildings, healthcare systems, and security. For example, since most sensor-based IoT devices still run on low-power batteries, the batteries need to be discharged. In addition, sensor-based IoT devices are often used in rough and unreachable areas for a long time without the possibility of charging or replacing the batteries. Therefore, energy constraint is the primary challenge in IoT networks. However, ensuring Quality of Service (QoS) requirements, including data transmission delays and high reliability issues, are other challenges of IoT. To overcome these challenges, data aggregation is the primary process of IoT. Data aggregation is described as the procedure that collects data from various sensors and IoT devices and then integrates it using an aggregation function and attempts to minimize traffic on the system network. Advantages of the data aggregation process ; These can be listed as minimizing traffic in the system, reducing delays in data transmission and improving energy consumption of IoT devices. There are some research papers on data aggregation in IoT. Some of these articles ignored mobile agent-based data aggregation mechanisms and focused only on client-server-based mechanisms. It is thought that the various review studies on existing client-server based data aggregation mechanisms do not meet the current and comprehensive needs of the existing literature (Yousefi et al., 2021). With this motivation, authors present a systematic literature review of both client-server-based and mobile agent-based data aggregation approaches. In this study, data aggregation is divided into two main categories : client-server based and mobile agent based. Client-server based data aggregation includes three groups : clustering-based, tree-based and centralized. The ultimate goal of the study is to provide a detailed comparison of data aggregation mechanisms, providing a road map for future studies.

Data aggregation is vital in wireless sensor network design as it reduces the number of packets that need to be transmitted by increasing energy efficiency as data is collected from nodes to the sink. With the emergence of the Internet of Things and machine-to-machine communication, new architectures and use cases are evolving. In addition, advances in edge and fog computing have also led to improvements in network architecture design. In fog computing, control, computation, and storage are moved to nodes close the network edge. In WSN, a single sink is generally used in aggregation algorithms. However, in architectures where fog computing is used, multiple sinks are needed to collect data. For example, in machine-to-machine communication scenarios, there may be actuators subscribing to sensors. In this case, data is not only aggregated and processed within the network, but also distributed to the actuators. As heterogeneity in IoT networks increases, the use of a single sink for data aggregation has been abandoned. Instead, there may be multiple gateways with connections located at the network edge, possibly operated by different parties. Most of the existing literature on data aggregation in WSNs is on single-sink scenarios. Therefore, it is necessary to develop solutions that combine both gateway selection and routing for data aggregation for optimal energy efficiency. (Fitzgerald et al., 2018) presents two new models for data aggregation in IoT. In the first model, called 1K, it is possible to aggregate data in any of the networks where there is more than one gateway. The second, called nK, is better applicable to scenarios where network actuators are used. This model allows joint data aggregation and distribution to all or a minimum number of multiple target nodes. These models are formulated as an optimization problem using mixed integer programming (MIP) for both wireless transmissions and the computation of the data aggregation function. While making the formulation, energy costs were taken into account and the minimum total energy target was achieved. In the study, network lifetime optimization was used as the key performance indicator metric. In addition, in this study, formulations for minimum-maximum energy optimization per node were provided for both models and their impact on network lifetime optimization was evaluated. In addition to sensor and gateway selection, routing subgraphs for data aggregation and distribution are also included in the optimization using models. In this study, a numerical study was carried out to compare the energy and solution times required for both models in different topologies and network scenarios with 10 to 40 nodes.

IoT includes sensors, actuators and devices that connect to the internet through different access networks, such as wireless sensor networks. Thanks to the wireless sensor network, there can be millions of connected devices in the network and they can work integrated for the same purpose. Devices, sensors and actuators in the wireless network can connect physical objects in the environment to the Internet, which allows access to physical objects anytime and anywhere. Sensors and devices in the wireless IoT network use sensing and communication technologies to work together and detect and access physical objects. In the IoT network, large amounts of data are transmitted from one point in the network to another. Currently, sensors are energy limited because they work with power such as batteries. In addition, sensors have to work actively for a long time in the physical environment, sometimes in conditions where battery replacement is not possible. This requires that the power they consume be used with maximum efficiency. It has been determined that the highest power consumption in IoT networks is during communication. At this stage, there are studies showing that if the data transmitted in packages is aggregated and transmitted to ensure energy efficiency, it consumes less power. It has been shown to be beneficial in terms of power consumption, bandwidth, load balancing, network lifetime and data accuracy in a wireless IoT network by using appropriate data aggregation methods.

In summary, the literature highlights the critical role of data aggregation in enhancing the efficiency and sustainability of IoT systems, particularly within energy-constrained wireless sensor networks. While various models and architectures—from centralized and cluster-based to mobile agent and fog-based approaches—have been proposed, many existing studies are still limited in addressing real-time, multi-type, high-frequency data from heterogeneous IoT sensors. This study aims to contribute to the field by proposing and evaluating a system capable of rapidly collecting, aggregating, and processing diverse sensor data in a centralized architecture, thereby addressing the limitations observed in current research.

2.2 Heterogeneous IoT networks and interoperability

As IoT technologies are becoming more dominant and even more WSNs are present in various different applications, there is a clear variety for IoT devices. Using multiple devices presents its own challenges, stemming from reasons such as difference in communication protocols, data schema differences etc. The variety in IoT environments present formidable challenges stemming from data heterogeneity, interoperability, and IoT device heterogeneity. These fundamental issues significantly impact the effectiveness of data processing, analysis, and the deployment of advanced technologies like Federated Learning (FL)(Samikwa et al., 2024).

Data heterogeneity in IoT refers to the highly diverse nature, structure, and statistical properties of data collected from various IoT devices and systems (Lin et al., 2025; Marshoodulla and Saha, 2023). This diversity arises from several key factors. IoT data originates from billions of distributed objects, including sensors, RFID readers, and other devices. This data is often found in varied formats such as XML or JSON (Marshoodulla and Saha, 2023). For instance, real-world digitalized building IoT data encompasses readings from a wide array of systems, including heating, ventilation, and air conditioning (HVAC) systems, lighting, photovoltaic panels, security, and electrical facilities (Lin et al., 2025). All these attributes of IoT networks results in diverse data sources and formats. Additionally, IoT networks demonstrate significant variability in data resolution, update frequency, and communication protocols. Sensors of the same type might operate at different sampling frequencies ; for example, some Voltage Sensors may capture data at 1-minute intervals, while others operate at 5-minute intervals, reflecting customized configurations for various use cases. Additionally, high-frequency data streams can contain gaps due to network latency or sensor downtime, resulting in missing data points. The data heterogeneity in this context poses several issues that are emphasized ; difficulty in deep-learning frameworks, time-series modeling complications, analytical biases, integration hurdles and data quality issues (Lin et al., 2025; Meng et al., 2025).

This study by (Lin et al., 2025) investigates the challenges of processing heterogeneous IoT data within deep-learning frameworks in digitalized buildings.

(Marshoodulla and Saha, 2023) focused on resolving the pervasive problem of data heterogeneity in the IoT, aiming to make sensor data accessible and interpretable across diverse systems and applications. Their core contribution was a Software Defined Network (SDN)-based IoT architectural framework that leverages semantic technologies for data integration and sharing. The researchers reported satisfactory results in converting heterogeneous data to homogeneous form in a real-time environment. The SDN-based system achieved an average throughput of 3.8 mbps, and SPARQL queries executed on the generated triples showed a satisfactory response time.

The research published by (Samikwa et al., 2024), tackles significant challenges in deploying FL for edge IoT environments, specifically addressing the dual issues of statistically heterogeneous training data and varied computing and communication resources across IoT devices. The researchers' core contribution was the introduction of Dynamic Federated Split Learning (DFL). They proposed this novel framework to boost training efficiency in dynamic, heterogeneous IoT settings through two main mechanisms : resource-aware split computing of deep neural networks, which involves dynamically allocating training tasks based on device capabilities and fluctuating communication ; and dynamic clustering of training participants, grouping devices by sub-model layer similarity to manage data heterogeneity without compromising privacy. To optimize the system, they formulated a multi-objective problem aiming to minimize global model training time, training loss, and total energy consumption on IoT devices, incorporating sensitivity parameters for application-specific trade-offs. They rigorously implemented a DFL prototype using Python and PyTorch, evaluating it on a real testbed of 10 diverse IoT devices like Jetson Nanos and Raspberry Pis. Power consumption was meticulously monitored, and experiments utilized CIFAR-10 and CINIC-10 datasets, simulating various non-IID data distributions. Compared to classic FL and Federated Split Learning (SFL), DFL demonstrated significant performance improvements. It reduced training time by up to 48%, consistently outperforming other methods across various power configurations and network throughputs. Furthermore, DFL achieved an impressive accuracy improvement of up to 32% by leveraging user clustering based on data similarities, effectively mitigating challenges from extreme non-IID data. Finally, DFL also led to substantial energy consumption reductions of up to 62.8%, strategically minimizing power usage by considering both computing and communication needs.

On the subject of FL on heterogeneous IoT devices, the study conducted by (Meng et al., 2025) describes a major challenge : performance degradation caused by extremely varied data and limited resources. The researchers propose an innovative FL algorithm that uses generative AI (GAI). Their core strategy involves leveraging diffusion models to create new data samples on individual IoT devices, effectively making local data distributions more uniform. This approach aims to reduce data heterogeneity and boost FL performance, even with minimal generated data. Beyond data synthesis, the study also incorporates resource scheduling to manage heterogeneous device capabilities. They formulated a joint optimization problem to minimize energy consumption while maximizing FL performance, subject to constraints like computational resources, communication bandwidth, and privacy. To solve this, they developed an efficient algorithm that optimizes data generation and communication resource allocation. Experimental validation with a central server and ten clients, simulating various data heterogeneity levels, proved their scheme's robustness and superiority. The results showed consistent convergence and superior performance, even where existing solutions failed, significantly reducing latency and energy consumption compared to other GAI-based methods. This validates its effectiveness for real-world applications by achieving lower error rates more rapidly and maintaining stable performance.

A systematic review by (Hiwale et al., 2023) investigates privacy-preserving methods for telemedicine, focusing on blockchain and federated learning. The primary goal of this study was to conduct a systematic literature review on previous research concerning privacy-preserving methods applied with blockchain and federated learning for telemedicine, aiming to design a secure, trustworthy, and accurate telemedicine model with a privacy guarantee. The authors acknowledge the challenges posed by data heterogeneity in healthcare, where cross-institute health data is crucial for developing robust global models, yet stringent regulations like HIPAA and GDPR restrict raw data sharing. FL is highlighted as a solution to this, enabling collaborative model training without transferring sensitive raw data, thereby accommodating horizontally and vertically distributed datasets. Furthermore, the review addresses interoperability as a key challenge for blockchain in healthcare, citing examples like SHAREChain and the use of FHIR standards to improve electronic health record (EHR) management and communication between disparate systems. The paper also considers IoT device heterogeneity, noting that technologies like the IoT and Internet of Medical Things (IoMT) are increasingly integrated into healthcare solutions to manage diverse sensor data and enhance privacy protection. The main contributions include a comprehensive review of relevant literature, a comparative analysis of privacy-preserving mechanisms used with blockchain and FL, and a discussion of their convergence for privacy-preservation in healthcare systems.

(Fusco and Aversano, 2020) introduce the Data Integration Framework (DIF), an approach for the semantic integration of heterogeneous data sources. Their central goal is to provide users with a unified view of data distributed across various information sources—whether structured, semi-structured, or unstructured—and to overcome challenges related to semantic heterogeneity and support interoperability with external systems. The authors highlight that traditional data integration efforts struggle with sources that employ different technologies, data representations, or design philosophies, often leading to redundant or incomplete information. DIF addresses this by using ontologies as a conceptual schema for both the individual data sources and the unified global view, distinguishing itself from previous systems that often relied on proprietary description languages which hindered broader communication. The framework acts as a generalization of both Global-as-View (GaV) and Local-as-View (LaV) approaches, leveraging data virtualization to ensure that users can query a conceptual domain model without needing to understand the underlying source details. The accomplishments include a detailed methodology for source wrapping, schema matching, schema merging, and query processing, all supported by a developed software prototype capable of integrating relational databases, spreadsheets, and XML data sources.

Another study conducted by (Houssein and Sayed, 2023) presents FedImpPSO, a boosted federated learning algorithm based on an improved Particle Swarm Optimization (PSO), specifically designed to enhance the robustness and efficiency of FL in unstable network conditions, particularly for healthcare IoT devices. The core objective of this research is to improve the accuracy of FL aggregation methods, which suffer significantly from the high volume of weights transmitted and received in erratic network situations. The paper underscores the inherent data heterogeneity in healthcare, where information from clinical institutions, patients, and various IoT devices are isolated and often in different formats, necessitating approaches that can handle non-IID (non-independent and identically distributed) and unbalanced datasets. A key challenge addressed is the IoT device heterogeneity, characterized by varying hardware capabilities, software configurations, and limited computational and communication resources of wearable and medical IoT devices. FedImpPSO aims to alleviate these limitations by minimizing data transfer overhead ; instead of full model weights, only the best and worst scores from local models are primarily transmitted to the server, with the best-scoring client submitting the full trained model. The authors demonstrate the effectiveness of FedImpPSO with accuracy improvements of 8.14% over Federated Average (FedAvg) and 2.5% over Federated PSO (FedPSO), validating its application in healthcare through case studies involving COVID-19 classification and cardiovascular disease prediction.

Another recent study by (He et al., 2024), introduces FedLGS (Federated Learning with Local Gradient Shielding), a new framework designed to improve personalized FL performance while protecting privacy on diverse, resource-constrained IoT devices. The researchers developed a dual-model approach for each device, featuring a shared model for global knowledge and a personalized one for local optimization, using local hypernetworks to balance collective intelligence with specific client data. Their most significant contribution is Local Gradient Shielding (LGS), a privacy method that conceals sensitive gradient information during local training, preventing data reconstruction and ensuring client privacy. The study also addressed data bias by reformulating the local risk function with a proximal term and scheduler. Through extensive experiments on medical datasets, FedLGS consistently outperformed nine state-of-the-art algorithms in terms of model performance and convergence across various data heterogeneity settings, confirming LGS's effectiveness in protecting client training data.

(Park et al., 2016) tackles key IoT challenges like data processing, resource management, network coexistence, and event detection by proposing innovative learning frameworks for self-organizing IoT devices. Uniquely, the researchers introduced a new framework based on cognitive hierarchy theory (CHT) to manage IoT heterogeneity. They comprehensively surveyed and classified three learning techniques applicable to IoT : Machine Learning for big data analytics, Sequential Learning (SL) for enhanced event detection, especially finite memory SL for resource-constrained devices, and Reinforcement Learning (RL) for resource management, noting its suitability for distributed systems due to computational simplicity. A central contribution was using CHT to model and manage IoT heterogeneity by mapping device resource capabilities to different levels of rationality. This allows each group of devices to select the most appropriate learning framework based on their available resources and application needs ; for instance, powerful smartphones might use complex RL, while simple sensors use finite memory SL. They applied CHT to practical problems like distributed uplink multiple access in heterogeneous networks. The research concluded that learning frameworks are vital for IoT's self-organization, and CHT effectively captures IoT heterogeneity, enabling the integration of diverse learning approaches tailored to different device types within a single system, with higher CHT rationality levels demanding more computational resources.

Interoperability in IoT is defined as the ability of diverse devices, systems, and platforms to seamlessly interact and exchange information, irrespective of their underlying protocols, technologies, or data formats. Achieving this is critical for unlocking the full economic potential of IoT applications. However, achieving this goal can pose some challenges because of several key issues. First, there's a significant variety of devices, and they all collect different kinds of data in various ways. Second, IoT setups often have their own unique ways of organizing data, making it tough to combine information. Even when using the same main tools, different networks and setups customize how data is shown, causing inconsistencies.

(Yacchirema and Palau, 2020), in their study, have tackled the significant problem of interoperability between diverse IoT devices and standard IoT platforms, which currently impedes efficient application development. The researchers proposed and implemented an architecture based on the oneM2M standard, featuring a core Interworking Proxy Entity (IPE) acting as middleware. The IPE was meticulously designed to perform protocol and message translation, fully mapping data models from heterogeneous IoT devices to the oneM2M standard's resource-based model. Integrated into a smart IoT gateway, the IPE incorporates various adapters for different communication technologies (e.g., ZigBee, Wi-Fi, LoRa) to abstract hardware/software differences and enable seamless interoperability. It standardizes data formats and maps translated messages to appropriate oneM2M resources, automatically propagating them to a cloud-based oneM2M server. For validation, the Smart IoT Gateway and IPE were implemented using Raspberry Pi and Python, while the oneM2M components were built with Eclipse OM2M. The architecture's feasibility and efficacy were confirmed through a practical use case in a port environment, specifically the INTER-LogP scenario, which involved integrating various sensors on a truck using diverse communication technologies. The findings conclusively demonstrated that this architecture enables robust interoperability between heterogeneous devices and oneM2M systems, facilitating data sharing with external platforms and remote device control. Preliminary results further indicated that information sharing via this architecture can significantly minimize queues, enhance control and security, regulate traffic flows, and optimize port operations, ultimately fostering global communication and interaction across IoT ecosystems.

2.3 IoT-based weather monitoring applications

In this (Kaya et al., 2023) study, authors introduce an intelligent anomaly detection approach aimed at minimizing weather forecasting problems that arise due to rapid urbanization and mass digitalization, thereby enhancing the accuracy and reliability of predictions. The core goal of their study is to make data produced by sensors more meaningful and manageable by filtering out missing, unnecessary, or anomalous data at the IoT edge, before it reaches the network layer. In terms of data acquisition, the study simulates a data stream from an IoT sensing layer, consisting of four-dimensional temperature, humidity, pressure, and time values. The raw data was generated using specific hardware components : an Arduino UNO R3, a BME280 sensor (for temperature, pressure, and humidity measurements), and a DS1302 clock module for real-time information. An ESP32 microcontroller facilitated the capture and transfer of this raw data. In the results, authors suggest that applying data filtering techniques at the IoT edge supports rapid and accurate decision-making at weather stations.

The research conducted by (Ali et al., 2024) present the design and development of an Internet of Things (IoT) based weather monitoring network specifically tailored for smart agriculture, with the objective of providing farmers with real-time, reliable, and localized information about weather conditions to mitigate risks affecting crops. A key contribution is a low-cost, hyper-local system that overcomes the limitations of traditional weather stations which often provide non-local or untimely data. For data acquisition, their system employs a network of solar-powered weather stations, each comprising a sensing unit, a power supply, and a MyFi gateway/router. The sensing unit is built around an ESP8266 microcontroller and incorporates several types of sensors, including a Davis 6410 Anemometer for wind speed and direction, a Davis 0.2 mm Smart Rain Gauge Sensor for rainfall, a Dallas DS18B20 for temperature, and a Bosch BME280 for relative humidity and atmospheric pressure. Data is wirelessly transmitted from these sensing units to a public MQTT broker (HiveMQ) via Wi-Fi, then subscribed by a local server for analysis and alert generation. The paper collects various types of weather parameters from different sensors and focuses on integrating them into a unified, localized monitoring solution. Another IoT application in weather monitoring systems is proposed by (Mohapatra and Subudhi, 2022), where they focus on the development and deployment of a cost-effective IoT platform for monitoring and archiving weather data in a residential area, utilizing open-source technologies. Their primary goal is to demonstrate the feasibility of building a scalable IoT weather station from scratch using open-source components, which offers advantages in terms of faster time-to-market, cost reduction, and enhanced privacy and safety by allowing inspection of the source code. In terms of data acquisition, the system's IoT nodes consist of a Raspberry Pi Zero W or a NodeMCU ESP8266 microcontroller, connected to a suite of sensors including a DHT22 humidity sensor, a BMP085 atmospheric pressure sensor (both capable of providing temperature readings), and a Sharp GP2Y1010AU0F dust sensor. The data collected from these sensors is then sent over the Internet to a remote Virtual Private Server (VPS), which runs the Thingsboard application as an MQTT broker to listen for and store the incoming telemetry data. It contributes by providing a methodological guide and proof-of-concept for building such a system from the ground up, ensuring data integrity through a well-defined architecture. In the study by (Bin Shahadat et al., 2020), researchers present an efficient IoT-based weather station system with the aim of providing readily accessible and up-to-date meteorological information for daily use. The central goal is to enable users to access weather updates for any location from anywhere via an internet connection. The paper's main contribution is the successful development and testing of a cost-effective and portable weather station based on the NodeMCU Board (utilizing the ESP8266 Wi-Fi microchip) and the Blynk – IoT application. For data acquisition, the system incorporates a DHT11 sensor for temperature and humidity, a BMP180 sensor for air pressure, and a SEN-00194 rain sensor. The ESP8266 module processes the raw sensor output and transmits it to the cloud using its Wi-Fi capabilities, where the Blynk-IoT application then retrieves and displays the information through virtual terminals. The study focuses on collecting different types of meteorological data from integrated sensors and comparing its accuracy against national weather information.

Furthermore, (Matuska et al., 2020) introduce a compact Internet of Things (IoT) module specifically designed for local weather monitoring. Their objective is to collect meteorological data for dual purposes : predicting power output from photovoltaic panels and evaluating air quality in inhabited areas. The primary accomplishment and contribution of their work is the design and development of this IoT module, which is built around an Arduino Uno microcontroller and an ESP8266 WiFi module for communication. For data acquisition, the module integrates a diverse set of sensors, including a DHT22 for humidity and temperature, a BMP180 for barometric pressure (also providing temperature), AS7262 and AS7263 sensors for visible and infrared light power density measurements respectively, and an MQ-135 air quality sensor capable of detecting various gases. Data from these sensors is transmitted to a remote web server using HTTP GET requests via the ESP8266 module. The paper implicitly handles data heterogeneity by integrating various types of environmental and meteorological sensor data (e.g., digital, analog, spectral, and air quality) into a unified system, and then storing them in a structured MySQL database for different analytical applications. The collected data, comprising diverse parameters, are visualized and made accessible through a web page, demonstrating the module's ability to serve multiple environmental monitoring needs.

On the topic of weather monitoring another important subject is correctly measuring wind related data. This data can be crucial for wind energy. To this end, the study by (Karvelis et al., 2020) introduces PortWeather. The proposed solution is a lightweight onboard solution for real-time wind direction and speed prediction, particularly designed for maritime applications where continuous internet connection is often unavailable. A significant contribution is the development of an online-learning machine learning algorithm that does not require a pre-training phase, allowing the system to continuously adapt to new data and remain functional irrespective of the vessel's changing geographical position. For data acquisition, the system integrates a commercial weather station, the Airmar 220WX, with an industrial IoT-edge data processing module, referred to as the "Marine Gateway". The Airmar station is capable of collecting true and apparent wind speed and direction, air temperature, barometric pressure, and relative humidity. Crucially, it incorporates an integrated GPS and compass system to calculate real wind parameters by compensating for boat movement. Data transfer from the weather station to the Marine Gateway is facilitated by the NMEA 2000 protocol. The Marine Gateway's core is a 32-bit ESP32 microcontroller, which performs all data computation, analysis, and forecasting locally on the edge device. While the paper does not explicitly detail data heterogeneity in terms of integrating from disparate systems or formats, it focuses on the challenges of processing complex, continuously changing local meteorological data, especially wind parameters, to provide reliable forecasts on a moving platform. The system's design enables it to handle the dynamic nature of marine environments without external dependencies, highlighting its robustness and practical utility.

(Hewage et al., 2020) aim to develop and evaluate a lightweight and novel short-to medium-range weather forecasting system for end-users, such as farmers, by leveraging time-series data from local weather stations and state-of-the-art machine learning techniques. Their core contribution is the application of Temporal Convolutional Neural (TCN) networks for localized weather prediction of up to 10 parameters, demonstrating its superior performance over Long Short-Term Memory (LSTM) and traditional machine learning methods. The system's goal is to provide fine-grained, location-specific, and accurate forecasts for up to 9 hours, directly addressing the limitations of regional or global forecasting models. For data acquisition, the system relies on six dedicated local weather stations deployed in various farms across the UK. These stations are designed as standalone units, featuring a Weatherboard for attaching various sensors (e.g., wind vane, anemometer, barometer, thermometer, hygrometer, rain gauge) to a Raspberry Pi device for local computation and data logging. The collected weather parameters are transmitted to a central data server at fine-grained temporal resolution (every 15 minutes) using a Global System for Mobile (GSM) module. The study explicitly addresses data heterogeneity by integrating 10 different "surface weather parameters" (Barometer, Air pressure, Temperature, Humidity, Wind speed, Wind direction, Rain rate, Rain, Dew point, and Heat index) into a unified dataset for model training and prediction, acknowledging that these inter-related parameters collectively contribute to accurate forecasts.

(Leelavinodhan et al., 2021) focus on designing and implementing an energy-efficient weather station for wind data collection, specifically to enable energy neutrality in remotely deployed IoT devices for precision agriculture. The primary goal is to optimize the energy consumption by intelligently managing the data collected and sent from sensors. Their significant contribution is the development and validation of a novel asynchronous optimization algorithm for wind data collection. This algorithm notably reduces the number of data transmissions, and consequently power consumption, by sending data only when the wind speed exceeds a user-defined threshold, deeming those data points "valuable". For data acquisition, their prototype weather station uses a Davis DS6410 cup anemometer to measure both wind speed and direction. The anemometer transduces wind direction via a linearly variable electrical resistance and wind speed by digital pulses from rotating cups. An Arduino Pro Mini microcontroller is employed to interface with the anemometer, using its internal Analog-to-Digital Converter (ADC) for direction and an external interrupt for speed pulses. Data transmission to a LoRaWAN gateway is handled by a Mini-Lora v1.1a communication module. While the study concentrates on optimizing the transmission of wind data (speed and direction) from a single type of sensor, it does not explicitly delve into data heterogeneity from multiple, disparate systems. Instead, its focus on process optimization for data transmission, based on the significance of the collected data, implicitly addresses challenges related to efficient resource usage for specialized IoT deployments. The study demonstrated potential to reduce the weather station's power consumption by more than 60% through this intelligent data filtering and asynchronous transmission approach.

These studies show that various issues in heterogeneity and interoperability exist in multiple studies. A comprehensive summary of the papers surveyed during this section can be found on table 2.1.

TABLE 2.1 – Overview of existing studies focusing on data aggregation, heterogeneity, interoperability and weather monitoring.

#	**Authors**	**Method**	**Main Focus**	**Key Findings**
1	(Hiwale et al., 2023)	Literature review on privacy-preserving IoT methods	Privacy-preserving methods using blockchain and federated learning.	Blockchain and FL enhance secure, trustworthy, accurate telemedicine with privacy.

#	Authors	Method	Main Focus	Key Findings
2	(Lin et al., 2025)	Benchmarking analysis and case study	To investigate IoT data heterogeneity challenges in digitalized buildings for predictive modeling and energy management	Multimodal data integration is critically needed for IoT analytics.
3	(Kaya et al., 2023)	Anomaly detection	To propose an intelligent anomaly detection approach for accurate and reliable weather forecasting at IoT edges.	Anomaly detection at IoT edge enhances data accuracy and trustworthiness.
4	(Fusco and Aversano, 2020)	Ontology-based data integration framework (DIF)	To present an ontology-based framework (DIF) for semantic integration of heterogeneous data sources.	DIF offers unified access to heterogeneous data through semantic integration.
5	(Marshoodulla and Saha, 2023)	Semantic technologies with SDN-based IoT architecture	SDN-based IoT framework utilizing semantics to resolve data heterogeneity	Framework effectively converts heterogeneous IoT data to semantically rich RDF, in real-time.
6	(Ali et al., 2024)	IoT-based weather monitoring system development	Surveys the role of IoT in agriculture for implementing smart farming, including weather monitoring.	IoT effectively enables smart farming through real-time weather monitoring.
7	(Houssein and Sayed, 2023)	Improved Particle Swarm Optimization (FedImpPSO)	To propose FedImpPSO to enhance Federated Learning accuracy and robustness in unstable network conditions for healthcare.	FedImpPSO improves FL accuracy for healthcare applications significantly.
8	(Wang et al., 2024)	Asymmetric Nash bargaining model	Enhancing accuracy and fairness of IoT data sharing by considering participant heterogeneity in electricity retail.	Model promotes efficient IoT data use, guides large-scale transactions.
9	(Samikwa et al., 2020)	Dynamic Federated Split Learning (DFL)	To address joint data and resource heterogeneity for distributed training in heterogeneous IoT environments.	DFL significantly improves training time, accuracy, and energy.
10	(Mohapatra and Subudhi, 2022)	IoT-based system development	Developing a cost-effective IoT-based weather monitoring system.	System uses MQTT, Raspberry Pi, NodeMcu for data collection.
11	(Bin Shahadat et al., 2020)	IoT-based weather station development	To establish an inexpensive and easy weather station system using wireless sensors and a microcontroller.	Processing weather parameters and transmitting data wirelessly.

#	Authors	Method	Main Focus	Key Findings
12	(Meng et al., 2025)	Generative AI-empowered FL with resource scheduling	To propose an FL algorithm using generative AI and resource scheduling to enhance performance on heterogeneous IoT devices.	Utilizes of a GAI model, reducing the impact of data heterogeneity.
13	(Matuska et al., 2020)	IoT module development	Description of the development of an IoT module for meteorological data collection for power output prediction in smart buildings.	IoT module collects temperature, humidity, pressure, air quality.
14	(Yacchirema and Palau, 2020)	Interworking architecture based on oneM2M	To propose an interworking architecture for seamless interoperability between heterogeneous IoT devices and oneM2M-based systems.	Architecture validates interworking via protocol translation and data mapping.
15	(Park et al., 2016)	Cognitive hierarchy theory for learning frameworks	An overview of emerging learning frameworks for IoT, focusing on resource constraints, heterogeneity, and QoS requirements.	CHT efficiently captures heterogeneous IoT devices and varying resource levels.
16	(Karvelis et al., 2020)	Machine learning (regression algorithms)	To develop a lightweight, onboard solution for real-time wind data prediction on microcontrollers.	Lightweight regression algorithms are efficient and reliable for wind prediction.
17	(He et al., 2024)	Personalized Federated Learning (pFL) with local gradient shielding (LGS)	To address data heterogeneity and privacy protection challenges in Federated Learning for resource-constrained IoT devices.	Proposed FedLGS framework improves model utility and privacy.
18	(Hewage et al., 2020)	Temporal Convolutional Neural (TCN) Network	To develop and evaluate a lightweight short-to medium-range weather forecasting system using modern AI technologies.	TCN network effectively predicts local weather parameters.
19	(Leelavinodhan et al., 2021)	Asynchronous optimization algorithm	Optimization of energy consumption for a weather station with wind data collection to achieve energy neutrality	Asynchronous optimization reduces data transmissions and minimizes energy consumption.
20	(Bellavista et al., 2017)	PRESS project architecture	Environmental sensing	Real-time environmental monitoring and congestion detection

#	**Authors**	**Method**	**Main Focus**	**Key Findings**
21	(Landaluce et al., 2020)	Integration of RFID and WSN	RFID and WSN technologies	Synergies between RFID and WSN for IoT applications
22	(Zhang et al., 2021)	IoT middleware requirements	Middleware for IoT	Interoperability, self-adaptation, scalability, reliability
23	(Khanna and Kaur, 2020)	Various IoT applications	IoT technology	Smart mobility, smart home, healthcare, agriculture, etc.
24	(Lohiya and Thakkar, 2020)	Basic structure of IoT technology	IoT applications	Health, transportation, security, energy efficiency, etc.
25	(Rahman et al., 2016)	Comparison of data collection techniques	NS-2 simulation	Evaluation of data aggregation techniques
26	(Gulati et al., 2022)	Conservation of energy in WSN	WSN and IoT energy consumption reducing methods.	Energy-efficient data collection techniques
27	(Rezaeibagha et al., 2020)	Secure IoT-based wearable monitoring	Cryptographic accumulator	Lightweight encryption for secure data collection
28	(Yousefi et al., 2021)	Literature review on data aggregation	Client-server vs. mobile agent-based	Comparison of aggregation approaches
29	(Fitzgerald et al., 2018)	Data aggregation in IoT	Optimization models	Energy-efficient data aggregation with multiple gateways

3 METHODOLOGY

(Zhou et al., 2016) describes IoT systems as a loosely coupled, decentralized, and dynamic system. In this system, "things" or everyday objects are interconnected and become active participants in various processes like business, logistics, information, and social interactions. Achieving full interoperability of interconnected devices is one of the main goals for IoT systems (Atzori et al., 2010). However, IoT systems consist of many distributed and interacting components that often differ in hardware, communication protocols, software interfaces, data formats, and semantics (Fortino et al., 2018). These attributes of the IoT ecosystem results in a wide range of solutions being utilized in different parts of IoT systems, in return makes it difficult to employ or develop full-fledged methodologies when facing these challenges. As a result of these conditions, facing data integration challenges is a common issue (Wang et al., 2012a).

This occurrence is a result of the heterogeneous nature of the IoT systems. Wide variety in communication protocols, semantics, interfaces and hardware forces developers to employ different methods to tackle this issue. In this study, we aim to build a middle-ware for integrating a variety of data sources, that differ in semantic and syntax. In the following sections of this chapter, the challenges of data heterogeneity and the methods used to address them are described.

3.1 Interoperability

These differences can cause difficulties when the data is integrated into a single data storage that decision makers will then use. This problem is even more prevalent when it comes to integration of IoT-sourced data, since IoT based systems generally use a variety of different sources to provide data. These sources can be different sensor devices, embedded systems, or any other micro-controller-based devices. Additionally, these systems can be fed with additional data sources (databases, APIs, etc.). (Xiao et al., 2014) explains interoperability as the interaction between IoT devices and users or other IoT devices and categorize interoperability in two categories ; syntactic interoperability and semantic interoperability.

Syntactic interoperability refers to the structure and syntax of data or the messages exchanged between devices (Hosseini and Dixon, 2016). Syntactic interoperability issues become more evident when the high number of communication protocols that are used in IoT systems are considered (Çorak et al., 2018). The need for standardization in IoT communication is a long time discussion (Saleem et al., 2018), but there are many challenges for defining clear standards when it comes to IoT devices. Among these challenges one of the primary reasons is the heterogeneous nature of the IoT devices. Many different devices of various specifications are interconnected in a IoT system and each device has their own needs when it comes to transferring and receiving messages. To give an example, a device with a large memory capacity can support more memory demanding protocols and receive or transmit bigger packages while a more lightweight protocol would be more beneficial for a device with less memory. Another significant limit when choosing a protocol is the battery capacity of the IoT device since it is very common for IoT devices to work on battery (Abbasian Dehkordi et al., 2020). The purpose of the device can also be a reason for differences between protocols (Çorak et al., 2018). As a result, variations in technical capabilities, power capacities, and operational purposes can individually or collectively lead to syntactic heterogeneity. If these differences are not addressed carefully, interoperability between devices can be significantly hindered, since changes in syntax and data transmission structure impact all aspects of communication.

On the other hand, semantic interoperability refers to a higher-level concept in heterogeneous systems. In this context, when devices interact, the exchanged messages and information between devices —or humans and devices— is correctly interpreted and semantically consistent (Xiao et al., 2014). Semantic modeling, often utilizing semantic Web technologies and ontologies, has become crucial for resolving the problem of interoperability in the IoT because of its distributed and heterogeneous nature (Wang et al., 2012b). Among the causes of issues in semantic interoperability is also the sheer amount of diversity between devices. This diversity originates from a lot of reasons ; difference in operational purposes, difference in providers, complexity of domain knowledge the device is located in, etc. (Wang et al., 2012b).

As explained by (Ferreira et al., 2022), interoperability is a key concept in IoT systems and should be ensured during their development. Therefore, issues resulting from heterogeneity must be identified and carefully managed, as both semantic and syntactic differences can lead to interoperability problems, creating significant challenges for developers (Xiao et al., 2014).

3.2 Data heterogeneity

As mentioned in the introduction of the chapter, one of the reasons of the heterogeneity of the data is the diversity of the data sources. IoT devices can vary in many qualities. This variety is not only a result of the general lack of standardization, but also a consequence of the purposes of the devices. In a survey study conducted by (Noaman et al., 2022), challenges caused by the heterogeneity of IoT devices are categorized into fourteen categories, and while two of these topics are directly related to data integration issue (heterogeneous data/data formats, integration of devices and data) four other categories are indirectly related to data heterogeneity (heterogeneity in standards,platform, management and configuration of devices, heterogeneity of devices issues, heterogeneous communication issues). Additionally, the difficulties of administrating a heterogeneous IoT systems should also be taken into consideration.

Based on the research by (Fortino et al., 2018), (Dave and Mittapally, 2024), and (Tu et al., 2020), data heterogeneity refers to the diverse nature of data encountered in various systems, particularly in the context of IoT, Industrial IoT, and smart buildings. More specifically, data heterogeneity in these environments includes :

— *Variety of sources :* Data originates from different distributed systems, applications, devices, and sensors. (Fortino et al., 2018) (Eneyew et al., 2022)

— *Different data types :* It can involve different types of data, such as IoT streaming data and static data. (Tu et al., 2020)

— *Diverse formats and structures :* Data arrives in various formats (e.g., .csv, .txt, .html, .xml, text, document, log) and structures (structured, semi-structured, non-structured, relational). (Chen et al., 2013)

— *Semantic differences :* Data may have different semantics or lack a shared interpretation, making it difficult to understand and use consistently. Multi-modal sensory data, for example, involves different modalities of data within a set. (Cheng et al., 2019) (Xiao et al., 2014)

— *Timing conflicts :* Data from different sources can arrive at different times or frequencies, leading to issues with aligning timestamps. (Tu et al., 2020)

— *Technology and Protocols :* Underlying systems and devices might utilize distinct communication protocols and technologies (e.g., Wi-Fi, Bluetooth, Zigbee, LoRaWAN). (Hamza et al., 2023) (Xiao et al., 2014)

Data heterogeneity introduces significant challenges that complicate the effective use and integration of data in modern systems. It makes it challenging to effectively combine data from disparate sources into a coherent, usable format (Tu et al., 2020). Moreover, the large volume, high variety, and rapid generation of data in IoT systems pose significant challenges for traditional databases and data management systems in handling and storing such data effectively (Hariri et al., 2019).

On the other hand, managing and storing data can be meaningless when there is a lack of semantic interoperability. As it was discussed in 3.1, semantic interoperability is essential for creating relations between gathered data and generating knowledge from this data. On top of that, a meaning-based data relation is crucial for a healthy user-device interaction. That's why the lack of semantic interoperability obstructs systems from fully utilizing relevant information and limits advanced functionalities like reasoning, learning, and decision-making in applications (Xiao et al., 2014).

3.3 Weather data

Weather refers to the daily fluctuations in atmospheric conditions. Data such as air pressure, temperature, wind speed and direction, humidity, and more which collected from various meteorological stations through sea, ground, radar, and other observations, reflect the current state of the weather. This data is then used for weather forecasting (Pandey et al., 2017). The practice of weather forecasting dates back to the nineteenth century (Bengtsson, 1980), and nowadays, it is an essential part of many people's lives (Leu et al., 2014). Moreover, weather forecasting is utilized intensively in many economic fields. Sectors such as agriculture, marine trade, air and ground transportation are some of the fields that can be named which are directly affected by weather condition, therefore strategic decisions are made using the forecasts (Jain and Jain, 2017).

Weather data can be obtained in different ways such as ground and sea stations, radars and satellites (Xiao et al., 2012). These sources can provide structured, semi-structured and unstructured data (Fathi et al., 2022). During different forecasting methods, data such as satellite images, sea-level temperature readings or atmospheric pressure can all be used separately or integrated. This attribute further enhances the need for data integration efforts as all the produced data can be invaluable.

In this study, the objective is to develop a weather data integration system that incorporates data from multiple weather stations into a unified storage platform, while also offering station administrators the ability to monitor and manage their stations. The objective of the integration is to build a weather data platform that will be accessible for both researchers and citizens. That is why ensuring the data was both up-to-date and available without delay was among the top priorities. The system should immediately include any new instance of data. Additionally, the integrity of the data is also an important factor. The data integrity implies avoiding duplicated, corrupted and missing data. And finally, the system should ensure the data consistency, which indicates the uniformity of data in measurement units and other ways of data representation.

Weather stations are capable of providing very detailed information. Varying from device to device, they can provide readings on temperature, pressure, wind speed, wind direction etc., the sampling rates and frequencies of all these readings depends on the type of sensor that is used on the station. Therefore, for a detailed dataset it is crucial to provide metadata for data that is represented. This metadata must be prepared and served for every different weather station. Also, the geospatial data and other identification data of the station must be included in this metadata management as this information is also crucial for administrators for maintaining their remote stations.

3.4 HTTP

HTTP (Hypertext Transfer Protocol) is a protocol used for structured communication within the Internet of Things. It follows a request/response model for client-server interactions, where the client initiates a request and the server provides a corresponding response (Wukkadada et al., 2018). The HTTP protocol sends multiple small packets to the server during connection establishment, which can result in high traffic, increased network resource usage, and potential network delays (Wukkadada et al., 2018), increasing the power consumption.

3.5 Web scraping

Data scraping, or web scraping, is the automated process of extracting information from websites or online platforms (Khder, 2021). It is commonly applied in areas like data analysis, research, marketing, and software development. Instead of copying data manually, scraping enables efficient and large-scale data collection from multiple web sources (Boegershausen et al., 2022). It is widely used by researchers (Vanden Broucke and Baesens, 2018), and also in domains like marketing (Boegershausen et al., 2022), e-commerce (Henrys, 2021), and tourism (Barbera et al., 2023).

A scraper is a program or script designed to interact with websites to extract their content (Khder, 2021). It begins by sending a request—typically via HTTP, the protocol that governs web data transfer. Websites are built using HTML, styled with CSS, and rendered as structured documents. The browser represents this content using the Document Object Model (DOM), where each page element is a node that can be accessed or modified programmatically.

The first step of scraping is locating the relevant pages and analyzing their DOM structures to identify where the desired data resides. CSS selectors or XPath are then used to target specific elements in the DOM (Vanden Broucke and Baesens, 2018). This usually involves sending an HTTP GET request, receiving the DOM, and then parsing it.

One popular tool for this task is Beautiful Soup (Richardson, 2024), a Python library for parsing HTML and XML. Used with the requests library (Reitz, 2025), it simplifies content retrieval and structures the page into a searchable parse tree, enabling efficient data extraction.

3.6 ETL for IoT

ETL stands for Extract, Transform, and Load, and is an integral part of modern applications that require faster business decisions based on a large number of data. Historically, ETL has been a fundamental software process within data warehouse architectures, used to integrate multiple heterogeneous and distributed data sources to provide centralized and unified access to data for decision support (Ali and Wrembel, 2017).

The process consists of three main phases :

— *Extraction :* This is the process of extracting data from multiple sources. These sources can be in different formats, including text files, videos, images, emails, social media, XML, JSON, CSV, and various database types (Ali and Wrembel, 2017).

— *Transformation :* In this phase, the extracted data is transformed to make it clean and consistent with the structure and requirements of the target system, such as a data warehouse or data lake (Mehmood and Anees, 2022). This involves applying a series of rules and functions to the data (Ali and Wrembel, 2017; Mehmood and Anees, 2022). Transformations can range from basic data substitution (e.g., replacing null values or changing "london" to "London, United Kingdom") to more advanced operations like data quality checks, cleaning data to remove errors and null values, standardizing values, integrating data into a common consistent dataset, removing duplicates, sorting, computing summaries, complex joins, row aggregation, or column splitting (Ali and Wrembel, 2017). This phase is also an opportunity to migrate data from one storage type to another. Data transformations are not necessarily single transactions and may require numerous steps to achieve the proper format.

— *Loading :* Loading is the final step where the processed (transformed) data is written into the target system — usually a data warehouse, database, or data lake — so it can be used for analysis, reporting, or other business purposes (Patil et al., 2011). Its importance stems from the core purpose of a data warehouse. A DW is designed to integrate multiple heterogeneous and distributed data sources to provide centralized and unified access to data for decision support (Ali and Wrembel, 2017). Therefore, the successful completion of the loading phase is crucial because if an ETL workflow fails to finish its execution within a specified time window, the data warehouse becomes outdated and cannot be utilized by its stakeholders. This means all the prior effort in extracting and transforming data would be in vain if the data cannot be successfully loaded and made available for analysis and decision-making (Ali and Wrembel, 2017).

In IoT systems, ETL processes are used to integrate heterogeneous data from diverse sources (Mehmood and Anees, 2022). Another benefit ETL processes provide in IoT systems is that ETL processes takes some of the computing responsibilities off of the IoT devices as transformation operations are done on ETL services.

3.6.1 Data sources

The data sources of the system can be divided into two categories. Primarily the data originates from the weather stations on site which produce weather data by using various sensors. An additional data source is the weather stations by Météo-France. The data from Météo-France can be accessed via a public API provided by French government. These stations have a bigger coverage but provide measurements that are less frequent. Data from different data sources are combined in one data storage system in their separate data schemas. These data are served to users, researchers and administrators.

In this section the data from these data sources and their properties are described. As mentioned before, the primary data source are the installed weather stations. These weather stations produce data and this data is collected by the developed cloud service. During extraction process, the first step is to correctly create the tables in-which the data will be stored. During that, the fields of the table must be defined with data types in mind. By correctly assigning data fields the integrity of the data is ensured. Additionally, another key point during design is the storage size of different data types.

Measurement	**Vantage Pro2**	**MetSens500**	**MeteoHelix IoT Pro**	**MeteoWind IoT Pro**
Temperature	X	X	X	
Barometric pressure	X	X	X	
Rainfall	X			
Rain rate	X			
Dew point	X		X	
Wind speed	X	X		X
Wind direction	X	X		X
Wind gust				X
Humidity	X	X	X	
Windchill	X			
Heat index	X			
Soil moisture	X			
Solar radiation	X	X	X	

TABLE 3.1 – Comparison of stations and their measurement capabilities.

4 PROPOSED APPROACH

In this chapter the proposed weather sensor data integration middleware will be discussed. As mentioned in 3, the heterogeneity of IoT devices demands correct data integration and administration strategies. During the design of the middleware certain points were key points. Primarily, data must be uniformed. That means the final form of the data must conform to the same format and must be represented in a single data source, accessible by stakeholders. Secondly, data must correctly represent the original semantics. This is significantly important for sensor data with different update intervals and units of measurement. And finally, any administrator or other stakeholder should be able to track and monitor the IoT devices. In the scope of this project, the distributed nature of the stations makes this goal crucial for the success of the system.

4.1 Weather stations

In this section, the weather stations that acted as IoT devices will be described in detail. There are four types of weather stations that are installed in the region, VantagePro2, MetSens500, MeteoHelix IoT Pro and MeteoWind IoT Pro.

4.1.1 Vantage Pro2

The Davis Vantage Pro2 is a professional-grade wireless or cabled weather station designed for accurate, real-time monitoring of weather conditions. It measures a wide range of parameters, including temperature, humidity, wind speed and direction, barometric pressure, and rainfall. Known for its reliability and durability, it is used in both personal and professional settings like agriculture, research, and education. The system is modular and can be expanded with additional sensors such as solar radiation or soil moisture sensors. It transmits data to a console or a data logger for further analysis or online sharing (Davis Instruments, 2025). The sensor accuracies and other details are given on table 4.1.

Measurement	Accuracy	Range	Update Interval
Temperature	±0.3°C	-40°C to 65°C	10 to 12 s
Barometric pressure	±0.8 mm Hg	410 to 820 mm Hg	1 min
Rainfall	±3% (up to 250 mm/hr)	0 to 999.8 mm (Daily)	20 to 24 s
Rain rate	±5% (intensities < 250 mm/h)	0 to 762 mm/h	20 to 24 s
Dew point	±1°C	-76°C to +54°C	10 to 12 s
Wind speed	±3.2 km/h	0 to 322 km/h	2.5 to 3 s
Wind direction	±3°	1 to 360°	2.5 to 3 s
Humidity	±2%	1–100 %HR	50 s to 1 min
Windchill	±1°C	-40°C to 65°C	10 to 12 s
Heat index	±1°C	-40°C to 74°C	10 to 12 s
Soil moisture	*Not specified*	0 to 200 cb	77 to 90 s
Solar radiation	±5%	0 to 1800 W/m²	50 s to 1 min

Table 4.1 – Measurement specifications of the Vantage Pro2.

4.1.2 MetSens500

The MetSens 500 by Campbell Scientific is a compact, all-in-one weather station designed to measure key meteorological parameters such as temperature, humidity, rainfall, wind speed and direction, and solar radiation (Campbell Scientific, Inc., 2025). It offers easy installation, reliable data logging, and durable performance for environmental monitoring, agriculture, and research applications. The details of the station are described on table 4.2.

Measurement	Accuracy	Calibration	Measuring Range
Temperature	±0.3 °C (at +20 °C)	0.1 °C	–40 – 70 °C
Barometric pressure	±0.5 hPa at 25 °C	0.1 hPa	300 – 1100 hPa
Rainfall	99% (< 120 mm/hr)	from 0.01 mm	0 – 1000 mm/hr
Wind speed	±3% (up to 40 m/s), ±5% (up to 60 m/s)	0.01 m/s	0.01 – 60 m/s
Wind direction	±3° up to 40 m/s, ±5° for 40 – 60 m/s	1°	0° – 359°
Humidity	±2% (at 20 °C, 10 – 90% RH)	1%	0 – 100%
Solar radiation	Annual drift ±1.5%	1 W/m²	0 – 2000 W/m²

Table 4.2 – Measurement specifications of the MetSens500.

4.1.3 MeteoHelix IoT Pro

The MeteoHelix IoT Pro is a compact, solar-powered micro-weather station designed by Barani Design Technologies. It offers meteorological measurements compliant with World Meteorological Organization (WMO) standards, making it suitable for applications ranging from agriculture to urban climate monitoring. It operates wirelessly using LoRaWAN, Sigfox, or NB-IoT communication protocols and is capable of transmitting data up to five times per hour. With a self-charging system, it can function for up to four months without sunlight, ensuring reliable operation in various environmental conditions (Barani Design, 2025a). Measurement capabilities of MeteoHelix is explained on table 4.3.

Measurement	Accuracy	Calibration	Measuring Range	Update Interval
Temperature	±0.1 °C (0 - 65 °C) ±0.2 °C (-45 - 65 °C)	0.01 °C	-45 °C - 65 °C	5 to 30 s
Barometric pressure	±0.5 hPa (0 - 55 °C)	0.02 hPa (mbar)	300 - 1100 hPa	0.1 s
Rainfall	±2%	0.1 - 0.5 mm (Rain gauge dependent)	4915 mm/hr at 0.2 mm resolution	Rain rates up to 600 mm/hr
Dew point	*Calculated*	0.1 °C	-45 °C - 65 °C	8 to 40 s
Humidity	±1.5 %RH (0 to 65 °C)	0.2 % RH	0 - 100 % RH	8 to 40 s
Solar radiation	5% of daily total	1 W/m²	0 - 2000 W/m²	$<$ 1 s

Table 4.3 – Measurement specifications of the MeteoHelix IoT Pro.

4.1.4 MeteoWind IoT Pro

The MeteoWind IoT Pro is a compact, wireless wind sensor designed to meet the highest World Meteorological Organization (WMO) standards. The sensor offers over four months of battery life without sunlight. It supports open communication protocols (Barani Design, 2025b). In table 4.4 the measurement details of the station are given.

Measurement	Accuracy	Calibration	Measuring Range	Update Interval
Wind speed	±0.2 m/s	0.1 m/s	0 – 85 m/s	4 Hz
Wind gust	*Not Specified*	1°	0 – 360°	4 Hz
Wind direction	2°	1°	0 – 360°	1 Hz

Table 4.4 – Measurement specifications for MeteoWind IoT Pro.

4.1.5 Météo-France weather stations

Météo-France is the official French meteorological service, responsible for weather forecasting, climate monitoring, and issuing weather warnings across mainland France and its overseas territories. Météo-France offers public access to meteorological data through its API portal, supporting transparency and enabling third-party applications (Météo-France, 2025). The shared data comes from the weather stations that are maintained by Météo-France and is accessed by using an API provided by Météo-France.

4.2 Integration of stations

During the integration step of the sensor data, there are two important aspects to take into consideration ; keeping the semantics of the data and providing uniformity. In this section the methods that were employed two achieve these goals are explained.

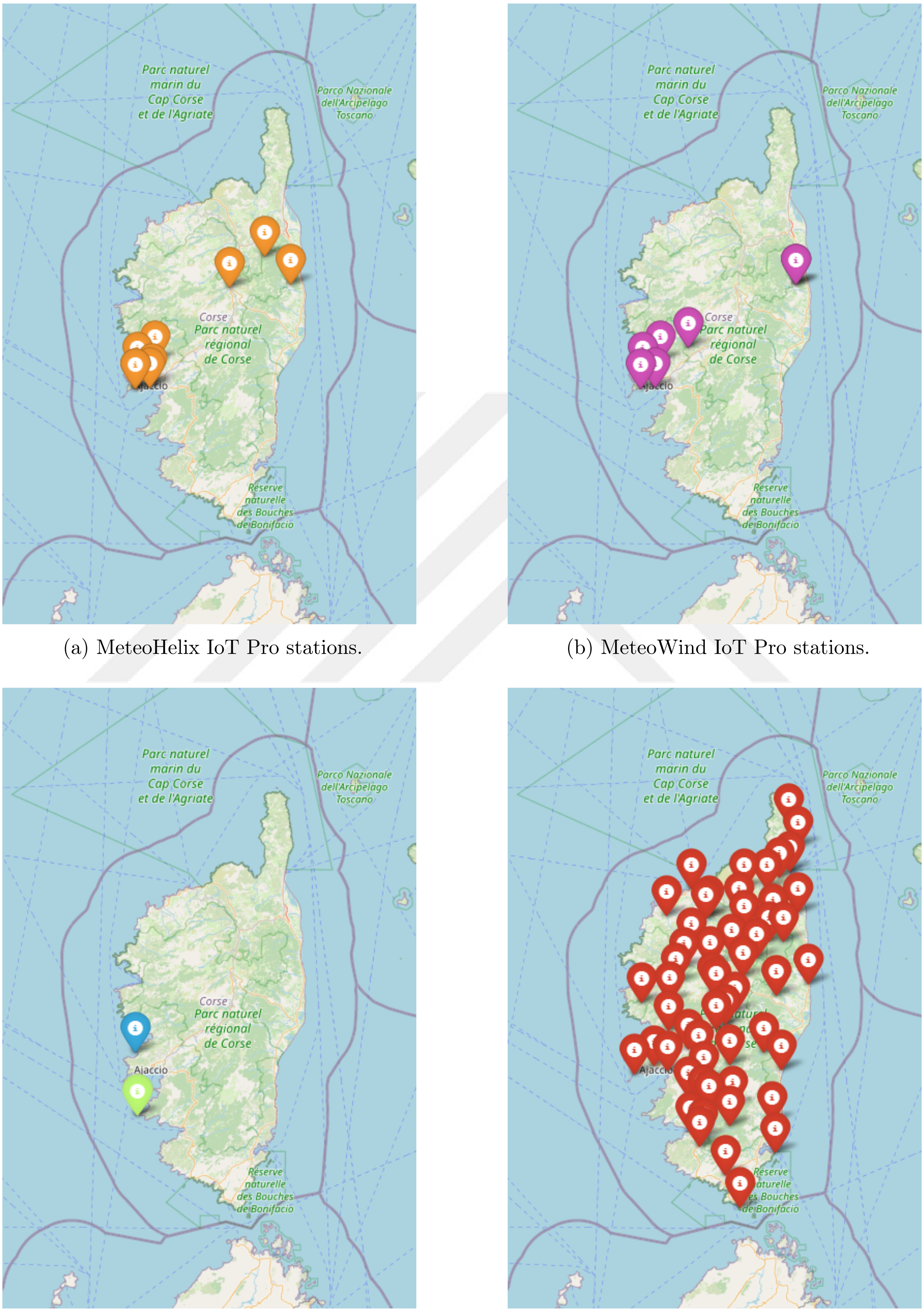
4.2.1 Building the database

In the integration step of the data, there are different criterion that should be taken into account when deciding on a data storage solution. The first step is to determine the structure and characteristics of the data. In the case of sensor-generated weather data there are some clear characteristics that we can detect.

Sensor data is continuous by nature. Unlike traditional data systems where data is produced by user interactions or actor actions, in a sensor-based IoT system data is generated at all times. With this in mind, it is safe to say that the data has a high velocity and volume.

Regarding the data format, the data generated by sensors is numerical and tabular. Although the structure of the data can change (e.g. every station has its own data structure) the weather stations do not generate image data or data in other formats. But everything that can be expected in a tabular format is possible in our IoT network (e.g. integer, float, string, boolean etc.).

Another important criteria is the needs of different stakeholders that will use the middleware. During the design of the middleware one of the goals was to propose a data storage for administrators of the system, researchers and users. Each of these stakeholders has different needs that should be met by the end result. For administration, the data storage should give insight on the current state of the network to the decision makers. This mechanism is crucial because the sparsely placed equipment is prone to errors and any device malfunctions should be assisted immediately. For researchers, the weather data is essential for weather prediction models and other studies that are conducted. Consequentially, data must have integrity and should have no errors in order to not impair the machine learning models that will be trained with the data. And finally, for end users the most important requirement is the availability of the system, the regular user should be able to access the system to view the current and



(a) MeteoHelix IoT Pro stations.

(b) MeteoWind IoT Pro stations.

(c) MetSens500 and Vantage Pro2 stations.

(d) Météo-France stations.

FIGURE 4.1 – Locations of stations.

past weather data in all the stations.

Evaluating these points, we can determine some key points. The data has high velocity, high volume and is in a singular format but in various schemas. When it comes to the end state of the data, it must correctly represent the current and past state of the IoT network while possessing integrity of the data for researchers' needs. And finally, the data should always be up to date and have consistency.

Two major options when choosing a data storage solution for a small to medium sized data volumes are SQL and NoSQL databases (Reetishwaree and Hurbungs, 2020). Both of these solutions have their advantages as well as the disadvantages. Relational databases are capable of providing up-to-date, consistent, secure and integrated data, and they assure these standards by complying to ACID rules. Although, these attributes also makes relational databases lack flexibility unlike their NoSQL counterparts (Khan et al., 2023).

NoSQL databases are more favorable when the data needs are unclear since they provide better potential to scale horizontally. Additionally, NoSQL databases are better at handling different data formats by being able to store different data formats in same structures. Availability of the NoSQL databases is also better than relational databases, but this comes at the expense of the consistency of the data (Rathore and Bagui, 2024).

The need for consistency, integrity and high quality in the final state of the data makes SQL options a favorable contender when choosing a data storage solution. Although for the end users' needs the availability of the storage is crucial, for other stakeholders the consistency and reliability of the system is more impactful. Besides, as discussed on the beginning of the section, the data format support of NoSQL databases provides no immediate advantages for the case in hand. The final argument that can be made for NoSQL databases is the option of scaling, however in the current state of the IoT network it seems unlikely that a need for growth will be needed in a recent time. Yet, it must be noted that a NoSQL solution would facilitate scaling better than a relational database solution.

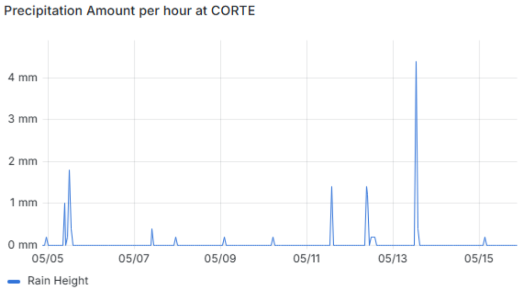
PostgreSQL is a powerful, open-source relational database management system known for its reliability, feature richness, and standards compliance. It supports advanced data types, complex queries, full ACID compliance, and extensibility through custom functions and plugins. Often used in enterprise and academic settings, PostgreSQL is ideal for handling large volumes of structured data securely and efficiently (PostgreSQL Global Development Group, 2024).

The entity relationship diagram of the relational database is represented in figure 4.3. In the figure, we can see that each sensor has its own table and data format. This design choice is made for several reasons.

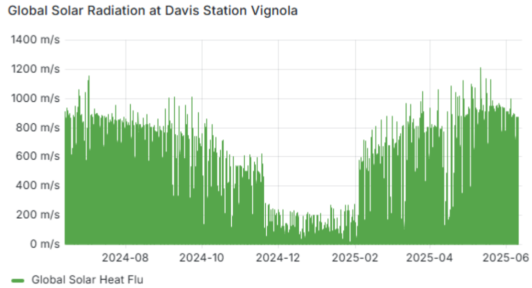
First, each sensor collects data at different time intervals. By storing them in separate tables, researchers can easily understand what kind of data each table contains. The heterogeneity of the data is more evident when table 3.1 is observed. As it can be seen on this table, the sensor data generated by the stations vary greatly depending on the source.

Second, even though the data is stored separately, these tables can still be queried together for visualization or data analysis purposes.

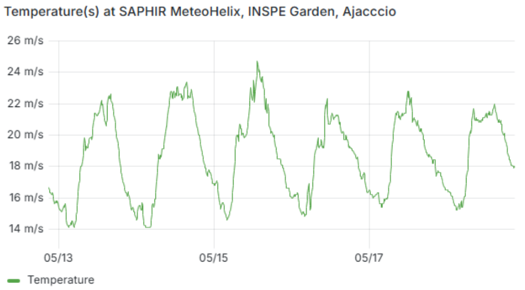
The table named 'station_metadata' stores detailed information about each station and its data structure. This serves two main purposes. For system administrators, this table is used to register each station and track its location. If there's a problem with the data, the metadata helps identify which station the issue came from by using foreign key relationships. For researchers, this table provides useful context about the data, such as the measurement frequency and units used by each station. It also explains some of the columns found in the station-specific tables, helping maintain the meaning and consistency of the data.



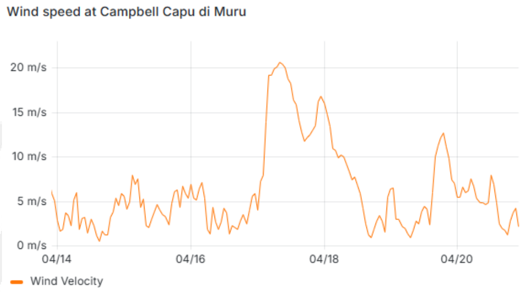
(a) Precipitation amount by MeteoFrance.



(b) Solar radiation readings by a VantagePro2 station.



(c) Temperature level readings from a Meteo-Helix station.



(d) Wind reading by MetSens500 station.

FIGURE 4.2 – Examples of station readings.

4.3 Middleware architecture

During the development of the API, FastAPI is used. FastAPI is a modern, fast (high-performance) web framework for building APIs with Python 3.7+ based on standard Python type hints. It is designed to be easy to use while enabling automatic validation, interactive documentation (with Swagger UI and ReDoc), and high speed thanks to its use of asynchronous programming. FastAPI is widely used in the development of machine learning, data processing, and microservices applications due to its strong performance and intuitive design (Ramírez, 2018).

In figure 4.4 the architecture of the middleware is visualized. MeteoHelix and MeteoWind stations periodically send HTTP messages to the API endpoint and data is extracted from their messages. For Vantage Pro2 and MétéoFrance stations, their API must be provoked periodically to receive the data as HTTP response. Unlike others, for MetSens500 the only way to access the data is the interface that is provided by the equipment's vendor. So, in set intervals a web scraper is invoked to scrape the website of the device's interface. And after the page is downloaded as HTML it is parsed inside the server and loaded to the database after appropriate transformations are applied.

This study encountered several noteworthy challenges related to the complexity and diversity of the data sources integrated into the system. Data were collected from three principal types of weather stations—MetSens 500, MeteoHelix and MeteoWind, and Vantage Pro2—with additional data provided by Météo-France. Each of these stations exhibited distinct technical characteristics, operated under different conditions, and utilized varied communication protocols. The stations were geographically dispersed, and the differences in environmental conditions, communication infrastructures, and power availability further increased the complexity of the integration process.

The heterogeneity among these stations manifested in several ways. The stations employed different types of sensors, measured overlapping yet not identical parameters, sampled data at varying frequencies, and transmitted data using distinct protocols. For example, the MeteoHelix and MeteoWind stations transmitted data via both HTTP and FTP protocols. These stations generated periodic HTTP GET requests directed at a predefined server endpoint, enabling regular data retrieval by the central system. Although this approach facilitated automated data collection, it necessitated careful management to handle potential network disruptions, latency, and transmission errors.

In contrast, the MetSens 500 stations presented a more complex case due to the absence of a dedicated programmatic interface. The manufacturer provided only a web-based interface designed primarily for human users rather than machine access. As a result, data from these stations were obtained through web scraping. This process involved periodic access to the web portal, extraction of the displayed HTML content, parsing of relevant data points, and conversion of the structured information into CSV format, which was then ingested by the ETL (Extract, Transform, Load) pipeline for processing and storage. While functional, this method introduced fragility into the system, as even minor changes in the website's structure or layout could disrupt data acquisition.

The Vantage Pro2 stations, by comparison, provided a direct API for sensor data access, which significantly simplified the integration process. However, the data were returned in semi-structured JSON format with nested fields of varying depth and complexity. Despite the availability of a data schema, the detailed parsing required to accurately extract and interpret all relevant data elements demanded meticulous implementation to avoid errors or omissions.

Beyond these technical integration challenges, the lack of uniformity in the data itself posed additional difficulties. One major issue was the inconsistency in time formats.

Stations reported timestamps using different formats and time zones, and in some cases, time zone metadata was omitted entirely. To ensure temporal consistency across the integrated dataset, a robust standardization process was implemented, converting all timestamps into a common format and unified time zone, typically Coordinated Universal Time (UTC). Such temporal alignment is essential in weather monitoring systems, where accurate timing directly affects the reliability of forecasts and decision support tools.

Another significant challenge arose from variations in units of measurement. Identical variables were reported in different units by different stations—for example, temperature in Celsius or Fahrenheit, and wind speed in meters per second, kilometers per hour, or miles per hour. A systematic unit harmonization process was applied using appropriate software tools to ensure that all values conformed to standard units, thereby enabling consistent analysis and interpretation.

Variations in data precision further complicated integration efforts. Differences in sensor quality and calibration were especially noticeable in parameters such as wind direction. Some stations provided wind direction data in cardinal points (e.g., N, NE, E), offering a coarse resolution limited to 16 discrete values, while others supplied continuous angular measurements with finer granularity. To preserve the integrity of analyses and visualizations, integration procedures were designed to align all data to the higher-precision standard wherever possible.

These challenges highlight the broader problem of heterogeneity in IoT-based weather monitoring systems. If not managed carefully, heterogeneity can result in data inconsistencies, integration delays, and errors that compromise the accuracy and timeliness of weather forecasts. Such shortcomings have serious implications, particularly in contexts where early warnings and precise predictions are critical to protecting life, property, and economic activity. As IoT deployments continue to expand across both urban and rural areas, the complexity associated with integrating diverse devices and systems is likely to grow. This underscores the urgent need for further research to develop standardized frameworks, tools, and best practices that can support the effective integration of heterogeneous IoT systems.

The middleware solution proposed in this study represents an initial step toward addressing these challenges. By supporting the collection, transformation, and integration of data from multiple heterogeneous sources, the middleware provides a foundation for building more reliable and responsive weather data systems. An example of the integrated data results produced by the middleware is presented in figure 4.2. Nevertheless, the long-term solution lies in the establishment of shared standards and interoperable architectures that can facilitate seamless data exchange and integration across diverse IoT platforms.

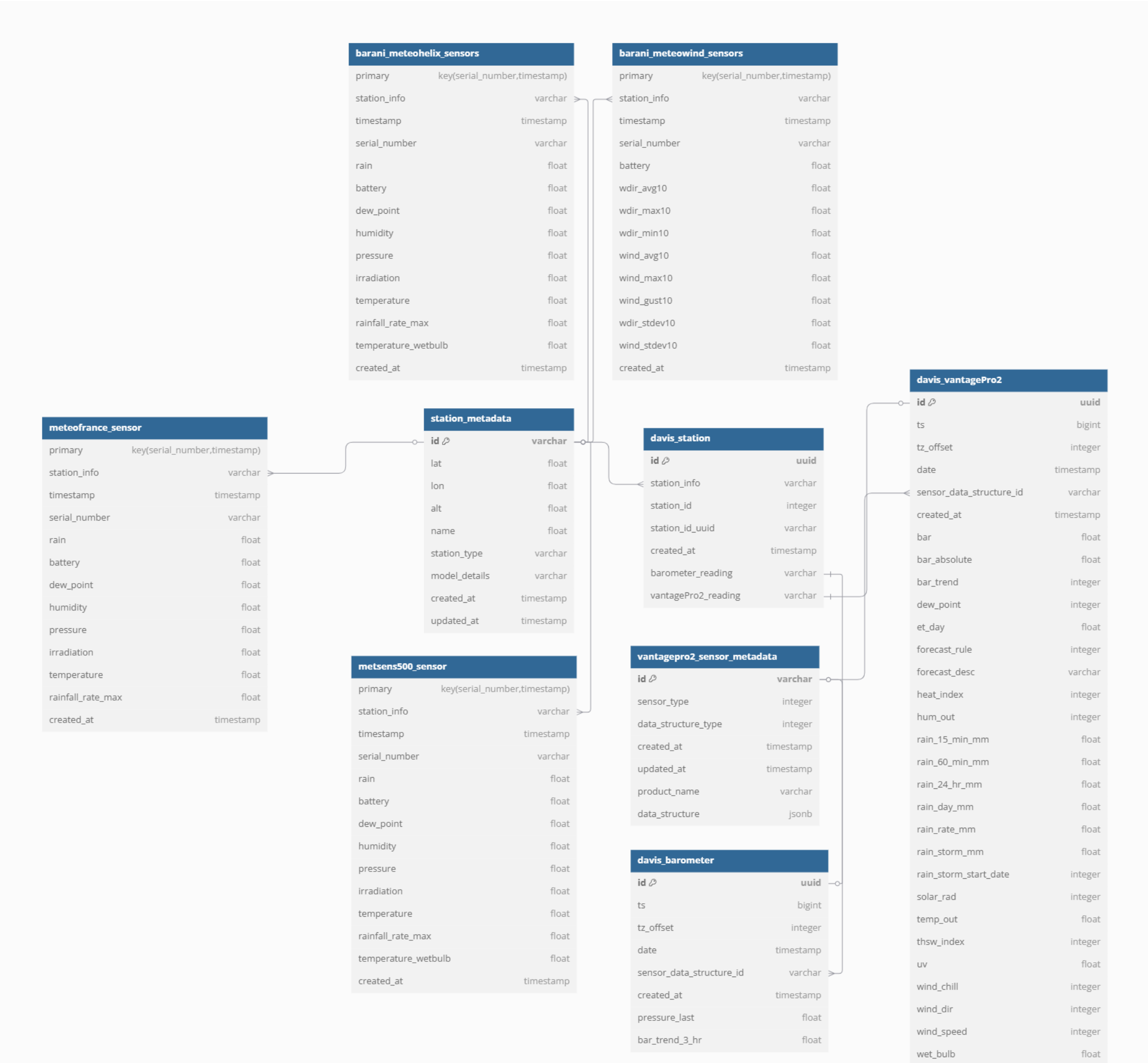


FIGURE 4.3 – Entity relationship diagram of the SQL database.

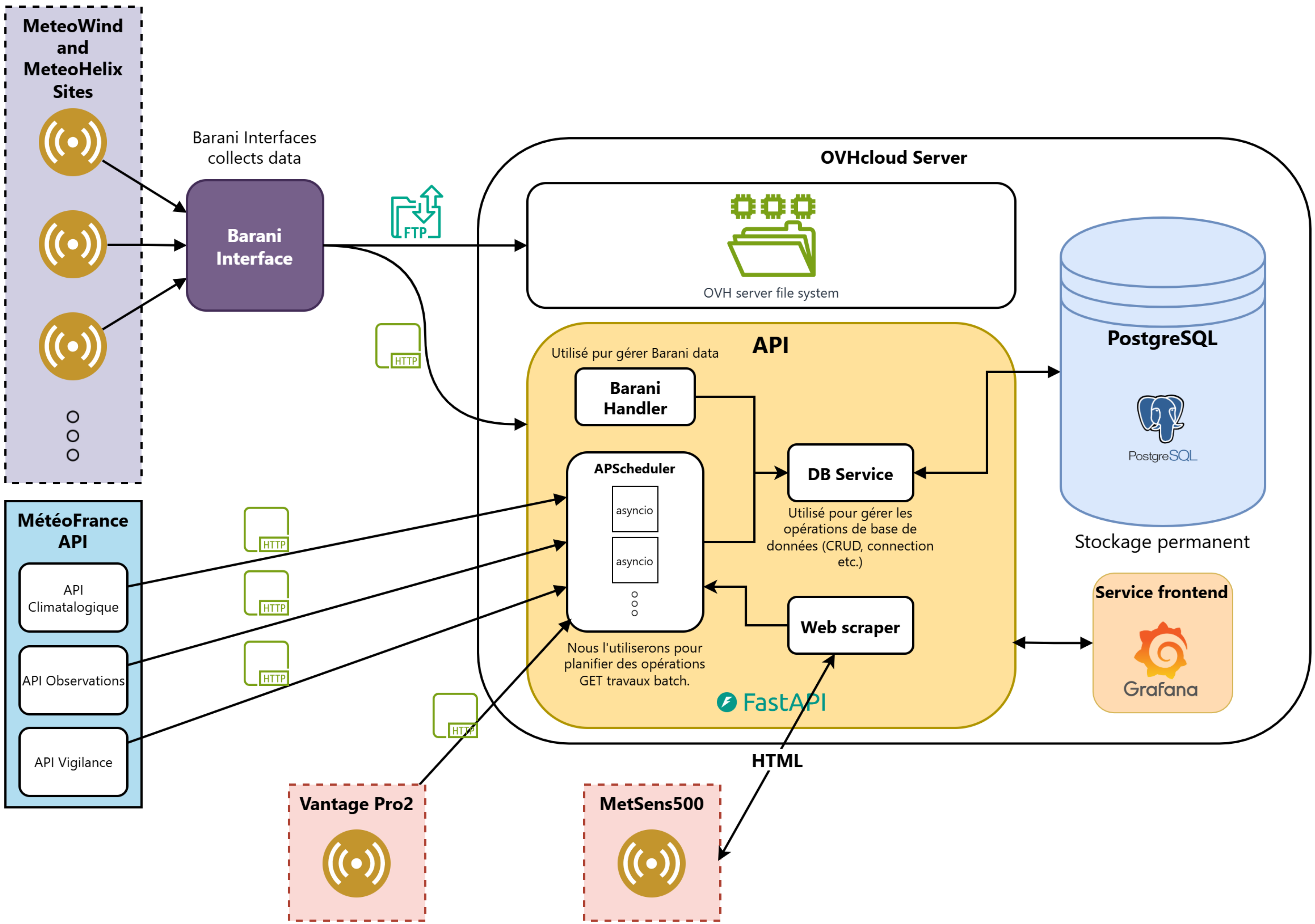


FIGURE 4.4 – Architecture of the middleware.

5 RESULTS

This section explores the methodology and presents the empirical results of performance evaluations conducted on the proposed middleware architecture.

Throughput testing was conducted to assess the system's capacity under simulated real-time data streaming conditions. This evaluation involved the concurrent ingestion of data from 20 simulated stations, each transmitting one data point per second, thereby mimicking a high-frequency sensor network. The simulation also aims to mimic a data load for 12 days. The observed performance metrics of the middleware under this sustained load are comprehensively presented in table 5.1. The table contains information on the average response time of the middleware, the storage size after the simulation and the amount of records that are collected.

Table 5.1 – Simulated performance metrics of middleware.

Metric Category	Metric Name	Units	Value	Description
Throughput	Max Data Rate	Messages/sec	20	Sustained rate under simulation
Latency	Avg. Response Time	ms	1127.62	Mean time for API response
Storage	Cumulative Rows	Records	1,495,080	Total records in database
Storage	Database Size	MB	194.3	Total storage occupied by data

Complementary to the simulated evaluations, the system's performance with actual collected data was also assessed. Figure 5.1 illustrates the cumulative data volume, represented by the row count, accumulated in the database over a 30-day period, utilizing real-world data acquired throughout the study.

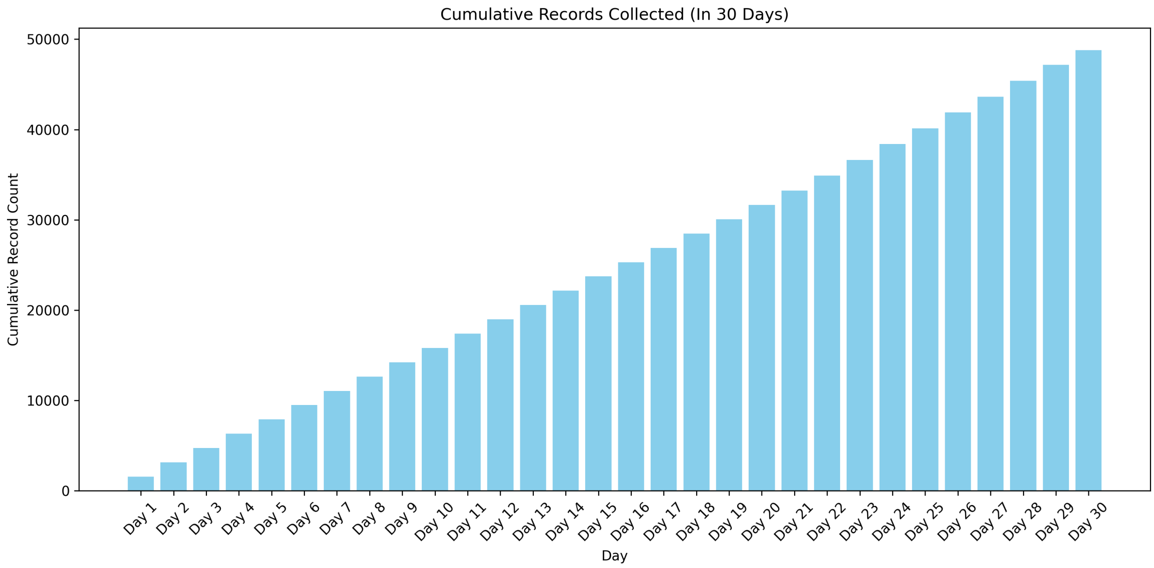


FIGURE 5.1 – Cumulative row count on the database over 30 days.

6 CONCLUSION AND DISCUSSION

In a world increasingly interconnected by billions of devices capable of sensing, communicating, and acting, the Internet of Things (IoT) holds remarkable promise for addressing some of humanity's most pressing challenges. From agriculture to healthcare, from urban mobility to environmental monitoring, IoT-based solutions enable data-driven decisions, automation, and efficiency at unprecedented scales. However, despite this potential, one of the most persistent and complex obstacles remains : heterogeneity in IoT networks and systems.

Heterogeneity refers to the diversity and incompatibility that arises from differences in hardware, software, communication protocols, data formats, energy profiles, and security mechanisms among IoT components. This issue stems largely from the lack of global standards in IoT device design and deployment, leading manufacturers and developers to adopt different technologies, often optimized for local or domain-specific needs rather than interoperability. As a result, heterogeneity manifests across almost every layer of IoT architecture — not only in data formats and transmission protocols but also in power sources, sensor precision, communication technologies (e.g., LoRaWAN, NB-IoT, WiFi, ZigBee), security schemes, and data processing methodologies.

Such diversity is not merely a technical nuisance ; it introduces significant risks. When IoT devices and systems cannot seamlessly communicate or integrate their data, critical applications — especially those that rely on real-time, high-accuracy data fusion — suffer. In domains like weather prediction, where multiple sensors, stations, and platforms must contribute harmonized data, heterogeneity can lead to delays, data loss, inconsistency, or misinterpretation of sensor readings. This, in turn, can degrade the accuracy of predictive models, potentially resulting in flawed forecasts that may undermine disaster preparedness, agricultural planning, and public safety. Yet, despite its centrality to IoT effectiveness, heterogeneity remains an under-researched challenge. There is a pressing need for more comprehensive studies, benchmarks, and frameworks that address how to mitigate its impact systematically, especially as IoT deployments scale globally.

In this study, we investigated the issues caused by heterogeneity in IoT networks, offering both theoretical analysis and practical experimentation through a real-time, field-based example. We focused particularly on rural environments, where heterogeneity is amplified by the use of varied device types — often dictated by availability, cost constraints, and harsh environmental conditions. To address these challenges, we proposed and implemented a middleware architecture specifically designed to facilitate data integration across heterogeneous IoT networks. This middleware is engineered to bridge the gap between diverse data producers and the varying needs of multiple stakeholders, including researchers, farmers, meteorological agencies, and policy makers — each of whom may have different requirements for data granularity, timeliness, and reliability.

The middleware presented in this study successfully collects, transforms, and integrates data from five distinct types of IoT-based data sources. These sources utilize different communication protocols and data transmission mechanisms, ranging from low-power wide-area networks to direct cellular connections. Through carefully designed scheduling strategies and a robust relational data model, the middleware ensures that incoming data streams are ingested in a timely manner and loaded into a PostgreSQL instance in the cloud after undergoing necessary transformations to ensure uniformity and integrity.

A critical component of the transformation process is the harmonization of data formats and semantics. Recognizing that each data source may employ different schemas, units of measurement, timestamp formats, and time zones, we implemented a multi-layered transformation strategy :

- Metadata management : The middleware maintains detailed metadata about the schemas, units, and formats of each source, allowing for dynamic and accurate transformation logic.
- Measurement unit standardization : Differences in units (e.g., Celsius vs. Fahrenheit, mm vs. inches for rainfall) are systematically reconciled to ensure consistent data representation across the integrated dataset.
- Timestamp normalization : All time-related fields are converted into a unified format and a single time zone, facilitating temporal alignment across sources.
- Temporal alignment of sensor readings : Our system accounts for the different sampling and transmission intervals of various sensors — information that is also captured in the metadata — to preserve the temporal fidelity of the data and avoid introducing artifacts during integration.

The impact of addressing heterogeneity effectively extends beyond data management ; it directly influences the quality of weather predictions and climate-related decisions.

Inaccurate, incomplete, or delayed integration of data from rural weather stations could lead to faulty forecasts, with severe consequences for communities dependent on timely warnings for floods, storms, or heatwaves. Conversely, robust integration frameworks like the one we propose can enhance forecast accuracy, enabling proactive responses and improving resilience against climate hazards.

Yet, as IoT networks continue to grow in scope and complexity, the problem of heterogeneity will only intensify. Rural deployments, in particular, are susceptible to this due to their reliance on diverse, often legacy, devices operating under strict power constraints and exposed to extreme environmental conditions. The lack of standardized deployment practices for these regions further complicates matters. Despite the vital importance of this topic, heterogeneity in IoT remains an underexplored field in academic and industrial research. There is an urgent need for deeper, more systematic investigations into how heterogeneity affects data integrity, system scalability, and predictive model accuracy, especially in high-stakes domains like meteorology.

In conclusion, this study contributes not only a practical solution in the form of middleware for heterogeneous IoT data integration but also highlights the critical importance of continued research into the risks and mitigation strategies associated with IoT heterogeneity. Future work must prioritize developing universal frameworks, standards, and toolkits to help system designers and operators navigate the complexities of integrating diverse IoT systems — ensuring that the full promise of IoT can be realized in service of society's most urgent needs, including accurate and timely weather prediction.

REFERENCES

Abbasian Dehkordi, S., Farajzadeh, K., Rezazadeh, J., Farahbakhsh, R., Sandrasegaran, K. and Abbasian Dehkordi, M. (2020). A survey on data aggregation techniques in iot sensor networks, *Wireless Networks* **26** : 1243–1263.

Alam, T. (2021). Cloud-based iot applications and their roles in smart cities, *Smart Cities* **4**(3) : 1196–1219.

Ali, H., Farooque, A. A., Abbas, F., Yaqub, R., Abdalla, A. and Soora, P. (2024). An iot based weather monitoring system for smart agriculture, *2024 IEEE Conference on Technologies for Sustainability (SusTech)*, pp. 378–382.

Ali, S. M. F. and Wrembel, R. (2017). From conceptual design to performance optimization of etl workflows : current state of research and open problems, *The VLDB Journal* **26**(6) : 777–801.

Almalki, F. A., Alsamhi, S. H., Sahal, R., Hassan, J., Hawbani, A., Rajput, N., Saif, A., Morgan, J. and Breslin, J. (2023). Green iot for eco-friendly and sustainable smart cities : future directions and opportunities, *Mobile Networks and Applications* **28**(1) : 178–202.

Ashraf, S. (2021). A proactive role of iot devices in building smart cities, *Internet of Things and Cyber-Physical Systems* **1** : 8–13.

Atitallah, S. B., Driss, M., Boulila, W. and Ghézala, H. B. (2020). Leveraging deep learning and iot big data analytics to support the smart cities development : Review and future directions, *Computer Science Review* **38** : 100303.

Atzori, L., Iera, A. and Morabito, G. (2010). The internet of things : A survey, *Computer Networks* **54**(15) : 2787–2805.
URL: *https ://www.sciencedirect.com/science/article/pii/S1389128610001568*

Awan, F. M., Saleem, Y., Minerva, R. and Crespi, N. (2020). A comparative analysis of machine/deep learning models for parking space availability prediction, *Sensors* **20**(1) : 322.

Barani Design (2025a). Meteohelix® iot pro weather station. Accessed : 2025-04-28.
URL: *https ://www.baranidesign.com/meteohelix-pro-weather-station*

Barani Design (2025b). Meteowind® iot pro. Accessed : 2025-04-28.
URL: *https ://www.baranidesign.com/meteowind-iot-pro*

Barbera, G., Araujo, L. and Fernandes, S. (2023). The value of web data scraping : An application to tripadvisor, *Big Data and Cognitive Computing* **7**(3) : 121.

Bauer, M., Sanchez, L. and Song, J. (2021). Iot-enabled smart cities : Evolution and outlook, *Sensors* **21**(13) : 4511.

Bellavista, P., Giannelli, C. and Zamagna, R. (2017). The pervasive environment sensing and sharing solution, *Sustainability* **9**(4) : 585.

Bengtsson, L. (1980). The weather forecast, *Pure and Applied Geophysics* **119**(3) : 515–537.
URL: *https ://doi.org/10.1007/BF00878157*

Bin Shahadat, A. S., Islam Ayon, S. and Khatun, M. R. (2020). Efficient iot based weather station, *2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)*, pp. 227–230.

Boegershausen, J., Datta, H., Borah, A. and Stephen, A. T. (2022). Fields of gold : Scraping web data for marketing insights, *Journal of Marketing* **86**(5) : 1–20.

Campbell Scientific, Inc. (2025). Metsens 500 weather station, `https://www.campbellsci.com/metsens500`. Accessed : 2025-04-22.

Chen, J., Chen, Y., Du, X., Li, C., Lu, J., Zhao, S. and Zhou, X. (2013). Big data challenge : a data management perspective, *Frontiers of computer Science* **7** : 157–164.

Cheng, S., Li, Y., Tian, Z., Cheng, W. and Cheng, X. (2019). A model for integrating heterogeneous sensory data in iot systems, *Computer Networks* **150** : 1–14.
URL: *https ://www.sciencedirect.com/science/article/pii/S1389128618312982*

Coromaldi, M. (2020). The impact of weather fluctuations and climate shocks on farmers' welfare : insights from rural ethiopia, *International Journal of Environmental Studies* **77**(4) : 619–635.
URL: *https ://doi.org/10.1080/00207233.2019.1695433*

Dave, D. M. K. and Mittapally, B. K. (2024). Data integration and interoperability in iot : Challenges, strategies and future direction, *Int. J. Comput. Eng. Technol.(IJCET)* **15** : 45–60.

Davis Instruments (2025). Vantage pro2. Accessed : 2025-05-01.
URL: *https ://www.davisinstruments.com/pages/vantage-pro2*

Deng, C., Han, M., Luo, J. and Zhang, J. (2025). The impact of weather shocks on rural credit cooperatives, *Finance Research Letters* **75** : 106872.
URL: *https ://www.sciencedirect.com/science/article/pii/S1544612325001370*

Eneyew, D. D., Capretz, M. A. M. and Bitsuamlak, G. T. (2022). Toward smart-building digital twins : Bim and iot data integration, *IEEE Access* **10** : 130487–130506.

Fathi, M., Haghi Kashani, M., Jameii, S. M. and Mahdipour, E. (2022). Big data analytics in weather forecasting : A systematic review, *Archives of Computational Methods in Engineering* **29**(2) : 1247–1275.
URL: *https ://doi.org/10.1007/s11831-021-09616-4*

Ferreira, L. C. B., Chaves, P. R., Assumpção, R. M., Branquinho, O. C., Fruett, F. and Cardieri, P. (2022). The three-phase methodology for iot project development, *Internet of Things* **20** : 100624.
URL: *https ://www.sciencedirect.com/science/article/pii/S2542660522001056*

Fitzgerald, E., Pióro, M. and Tomaszwski, A. (2018). Energy-optimal data aggregation and dissemination for the internet of things, *IEEE Internet of Things Journal* **5**(2) : 955–969.

Fortino, G., Russo, W., Savaglio, C., Shen, W. and Zhou, M. (2018). Agent-oriented cooperative smart objects : From iot system design to implementation, *IEEE Transactions on Systems, Man, and Cybernetics : Systems* **48**(11) : 1939–1956.

Fusco, G. and Aversano, L. (2020). An approach for semantic integration of heterogeneous data sources, *PeerJ Computer Science* **6** : e254.

Gao, J. and Mills, B. F. (2018). Weather shocks, coping strategies, and consumption dynamics in rural ethiopia, *World Development* **101** : 268–283.
URL: *https ://www.sciencedirect.com/science/article/pii/S0305750X17302978*

Gulati, K., Boddu, R. S. K., Kapila, D., Bangare, S. L., Chandnani, N. and Saravanan, G. (2022). A review paper on wireless sensor network techniques in internet of things (iot), *Materials Today : Proceedings* **51** : 161–165.

Hamza, A. S., Tashakkori, R., Underwood, B., O'Brien, W. and Campell, C. (2023). Beelive : The iot platform of beemon monitoring and alerting system for beehives, *Smart Agricultural Technology* **6** : 100331.
URL: *https ://www.sciencedirect.com/science/article/pii/S2772375523001600*

Hariri, R. H., Fredericks, E. M. and Bowers, K. M. (2019). Uncertainty in big data analytics : survey, opportunities, and challenges, *Journal of Big data* **6**(1) : 1–16.

He, Z., Zhang, F., Li, Y., Cao, Y. and Cai, Z. (2024). Privacy-enhanced personalized federated learning with layer-wise gradient shielding on heterogeneous iot data, *IEEE Internet of Things Journal* **11**(15) : 25771–25781.

Henrys, K. (2021). Importance of web scraping in e-commerce and e-marketing, *Available at SSRN 3769593* .

Hewage, P., Behera, A., Trovati, M., Pereira, E., Ghahremani, M., Palmieri, F. and Liu, Y. (2020). Temporal convolutional neural (tcn) network for an effective weather forecasting using time-series data from the local weather station, *Soft Computing* **24** : 16453–16482.

Hiwale, M., Walambe, R., Potdar, V. and Kotecha, K. (2023). A systematic review of privacy-preserving methods deployed with blockchain and federated learning for the telemedicine, *Healthcare Analytics* **3** : 100192.
URL: *https ://www.sciencedirect.com/science/article/pii/S277244252300059X*

Hosseini, M. and Dixon, B. E. (2016). Chapter 8 - syntactic interoperability and the role of standards, *in* B. E. Dixon (ed.), *Health Information Exchange*, Academic Press, pp. 123–136.
URL: *https ://www.sciencedirect.com/science/article/pii/B9780128031353000086*

Houssein, E. H. and Sayed, A. (2023). Boosted federated learning based on improved particle swarm optimization for healthcare iot devices, *Computers in Biology and Medicine* **163** : 107195.

Hussain, A., Memon, J. A. and Hanif, S. (2020). Weather shocks, coping strategies and farmers' income : A case of rural areas of district multan, punjab, *Weather and Climate Extremes* **30** : 100288.

Intergovernmental Panel On Climate Change (2007). Climate change 2007 : the physical science basis, *Agenda* **6**(07) : 333.

Jain, H. and Jain, R. (2017). Big data in weather forecasting : Applications and challenges, *2017 International conference on big data analytics and computational intelligence (ICBDAC)*, IEEE, pp. 138–142.

Karvelis, P., Mazzei, D., Biviano, M. and Stylios, C. (2020). Portweather : A lightweight onboard solution for real-time weather prediction, *Sensors* **20**(11) : 3181.

Kaya, Ş. M., İşler, B., Abu-Mahfouz, A. M., Rasheed, J. and AlShammari, A. (2023). An intelligent anomaly detection approach for accurate and reliable weather forecasting at iot edges : A case study, *Sensors* **23**(5) : 2426.

Khan, W., Kumar, T., Zhang, C., Raj, K., Roy, A. M. and Luo, B. (2023). Sql and nosql database software architecture performance analysis and assessments—a systematic literature review, *Big Data and Cognitive Computing* **7**(2) : 97.

Khanna, A. and Kaur, S. (2020). Internet of things (iot), applications and challenges : a comprehensive review, *Wireless Personal Communications* **114** : 1687–1762.

Khder, M. A. (2021). Web scraping or web crawling : State of art, techniques, approaches and application., *International Journal of Advances in Soft Computing & Its Applications* **13**(3).

Landaluce, H., Arjona, L., Perallos, A., Falcone, F., Angulo, I. and Muralter, F. (2020). A review of iot sensing applications and challenges using rfid and wireless sensor networks, *Sensors* **20**(9) : 2495.

Leelavinodhan, P. B., Vecchio, M., Antonelli, F., Maestrini, A. and Brunelli, D. (2021). Design and implementation of an energy-efficient weather station for wind data collection, *Sensors* **21**(11) : 3831.

Leu, J.-S., Su, K.-W. and Chen, C.-T. (2014). Ambient mesoscale weather forecasting system featuring mobile augmented reality, *Multimedia tools and applications* **72** : 1585–1609.

Lin, X., Prabowo, A., Razzak, I., Xue, H., Amos, M., Behrens, S. and Salim, F. D. (2025). A gap in time : The challenge of processing heterogeneous iot data in digitalized buildings, *IEEE Pervasive Computing* pp. 1–13.

Lohiya, R. and Thakkar, A. (2020). Application domains, evaluation data sets, and research challenges of iot : A systematic review, *IEEE Internet of Things Journal* **8**(11) : 8774–8798.

Marshoodulla, S. Z. and Saha, G. (2023). An approach towards removal of data heterogeneity in sdn-based iot framework, *Internet of Things* **22** : 100763.
URL: *https ://www.sciencedirect.com/science/article/pii/S2542660523000860*

Matuska, S., Hudec, R. and Sinko, M. (2020). Internet of things for local weather monitoring, *2020 ELEKTRO*, pp. 1–4.

Mehmood, E. and Anees, T. (2022). Distributed real-time etl architecture for unstructured big data, *Knowledge and information systems* **64**(12) : 3419–3445.

Mendelsohn, R., Basist, A., Kurukulasuriya, P. and Dinar, A. (2007). Climate and rural income, *Climatic Change* **81**(1) : 101–118.

Meng, Z., Li, Z., Hou, X., Xu, M., Xia, Y., Zhang, Z. and Song, S. (2025). Enhancing federated learning performance on heterogeneous iot devices using generative artificial intelligence with resource scheduling, *IEEE Internet of Things Journal* **12**(10) : 13286–13296.

Mohapatra, D. and Subudhi, B. (2022). Development of a cost-effective iot-based weather monitoring system, *IEEE Consumer Electronics Magazine* **11**(5) : 81–86.

Météo-France (2025). Météo-France Public Data Portal, `https://donneespubliques.meteofrance.fr/`. Accessed : 2025-05-03.
URL: *https ://donneespubliques.meteofrance.fr/*

Nguyen, T. T., Nguyen, T.-T., Le, V.-H., Managi, S. and Grote, U. (2020). Reported weather shocks and rural household welfare : Evidence from panel data in northeast thailand and central vietnam, *Weather and Climate Extremes* **30** : 100286.

Noaman, M., Khan, M. S., Abrar, M. F., Ali, S., Alvi, A. and Saleem, M. A. (2022). Challenges in integration of heterogeneous internet of things, *Scientific Programming* **2022**(1) : 8626882.
URL: *https ://onlinelibrary.wiley.com/doi/abs/10.1155/2022/8626882*

Pandey, A. K., Agrawal, C. P. and Agrawal, M. (2017). A hadoop-based weather prediction model for classification of weather data, *2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, IEEE, pp. 1–5.

Park, T., Abuzainab, N. and Saad, W. (2016). Learning how to communicate in the internet of things : Finite resources and heterogeneity, *IEEE Access* **4** : 7063–7073.

Patil, P. S., Rao, S. and Patil, S. B. (2011). Data integration problem of structural and semantic heterogeneity : data warehousing framework models for the optimization of the etl processes, *Proceedings of the International Conference & Workshop on Emerging Trends in Technology*, ICWET '11, Association for Computing Machinery, New York, NY, USA, p. 500–504.
URL: *https ://doi.org/10.1145/1980022.1980130*

PostgreSQL Global Development Group (2024). PostgreSQL : About, `https://www.postgresql.org/about/`. Accessed : 2025-05-23.

Pourghebleh, B. and Navimipour, N. J. (2017). Data aggregation mechanisms in the internet of things : A systematic review of the literature and recommendations for future research, *Journal of Network and Computer Applications* **97** : 23–34.

Rahman, H., Ahmed, N. and Hussain, I. (2016). Comparison of data aggregation techniques in internet of things (iot), *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, IEEE, pp. 1296–1300.

Ramírez, S. (2018). Fastapi : Fastapi framework, high performance, easy to learn, fast to code, ready for production. Accessed : 2025-05-23.
URL: *https ://fastapi.tiangolo.com/*

Rathore, M. and Bagui, S. S. (2024). Mongodb : Meeting the dynamic needs of modern applications, *Encyclopedia* **4**(4) : 1433–1453.

Reetishwaree, S. and Hurbungs, V. (2020). Evaluating the performance of sql and nosql databases in an iot environment, *2020 3rd International Conference on Emerging Trends in Electrical, Electronic and Communications Engineering (ELECOM)*, pp. 229–234.

Reitz, K. (2025). Requests : Http for humans, `https://requests.readthedocs.io/en/latest/`. Accessed : 2025-03-11.

Rezaeibagha, F., Mu, Y., Huang, K. and Chen, L. (2020). Secure and efficient data aggregation for iot monitoring systems, *IEEE Internet of Things Journal* **8**(10) : 8056–8063.

Richardson, L. (2024). Beautiful soup documentation. Accessed : 2025-03-8.
URL: *https ://www.crummy.com/software/BeautifulSoup/bs4/doc/*

Saleem, J., Hammoudeh, M., Raza, U., Adebisi, B. and Ande, R. (2018). Iot standardisation : challenges, perspectives and solution, *Proceedings of the 2nd International Conference on Future Networks and Distributed Systems*, ICFNDS '18, Association for Computing Machinery, New York, NY, USA.
URL: *https ://doi.org/10.1145/3231053.3231103*

Samikwa, E., Di Maio, A. and Braun, T. (2024). Dfl : Dynamic federated split learning in heterogeneous iot, *IEEE Transactions on Machine Learning in Communications and Networking* **2** : 733–752.

Samikwa, E., Voigt, T. and Eriksson, J. (2020). Flood prediction using iot and artificial neural networks with edge computing, *2020 International Conferences on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics)*, IEEE, pp. 234–240.

Tu, D. Q., Kayes, A., Rahayu, W. and Nguyen, K. (2020). Iot streaming data integration from multiple sources, *Computing* **102**(10) : 2299–2329.

Vanden Broucke, S. and Baesens, B. (2018). *Practical Web scraping for data science*, Springer.

Wang, B., Guo, Q., Xia, T., Li, Q., Liu, D. and Zhao, F. (2024). Cooperative iot data sharing with heterogeneity of participants based on electricity retail, *IEEE Internet of Things Journal* **11**(3) : 4956–4970.

Wang, W., De, S., Toenjes, R., Reetz, E. and Moessner, K. (2012a). A comprehensive ontology for knowledge representation in the internet of things, *2012 IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications*, pp. 1793–1798.

Wang, W., De, S., Toenjes, R., Reetz, E. and Moessner, K. (2012b). A comprehensive ontology for knowledge representation in the internet of things, *2012 IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications*, pp. 1793–1798.

Wukkadada, B., Wankhede, K., Nambiar, R. and Nair, A. (2018). Comparison with http and mqtt in internet of things (iot), *2018 International Conference on Inventive Research in Computing Applications (ICIRCA)*, pp. 249–253.

Xiao, G., Guo, J., Xu, L. D. and Gong, Z. (2014). User interoperability with heterogeneous iot devices through transformation, *IEEE Transactions on Industrial Informatics* **10**(2) : 1486–1496.

Xiao, Z., Liu, B., Liu, H. and Zhang, D. (2012). Progress in climate prediction and weather forecast operations in china, *Advances in Atmospheric Sciences* **29**(5) : 943–957.
URL: *https ://doi.org/10.1007/s00376-012-1194-9*

Yacchirema, D. and Palau, C. (2020). Interworking of onem2m-based iot systems and heterogeneous iot devices, *2020 XLVI Latin American Computing Conference (CLEI)*, pp. 262–267.

Yousefi, S., Karimipour, H. and Derakhshan, F. (2021). Data aggregation mechanisms on the internet of things : a systematic literature review, *Internet of Things* **15** : 100427.

Zhang, X., Manogaran, G. and Muthu, B. (2021). Iot enabled integrated system for green energy into smart cities, *Sustainable Energy Technologies and Assessments* **46** : 101208.

Zhou, M., Fortino, G., Shen, W., Mitsugi, J., Jobin, J. and Bhattacharyya, R. (2016). Guest editorial special section on advances and applications of internet of things for smart automated systems, *IEEE Transactions on Automation Science and Engineering* **13**(3) : 1225–1229.

Çorak, B. H., Okay, F. Y., Güzel, M., Murt, and Ozdemir, S. (2018). Comparative analysis of iot communication protocols, *2018 International Symposium on Networks, Computers and Communications (ISNCC)*, pp. 1–6.