

BigData: Maritime Data Integration and Analytics

A Comprehensive Approach to Real-Time Vessel Tracking, Historical Analysis, and Environmental Monitoring

Paweł Pozorski, Paweł Florek, Hubert Sobociński
November 2025

Contents

1	Introduction	2
2	Project Aim	3
3	Stakeholders	3
4	Data Sources	4
4.1	AIS Stream API	4
4.2	Maronesia API	4
4.3	Open Meteo API	4
5	Separation of Work	5
6	System Architecture	5
7	Example Analytical Use Cases	6

1 Introduction

Maritime operations generate vast amounts of data from vessel movements, ports, and environmental conditions. Efficiently collecting, processing, and analyzing this data is critical for ensuring safety, optimizing logistics, and supporting research and regulatory oversight.

This project aims to build a comprehensive maritime data platform that integrates real-time AIS streams, historical vessel and port information, and environmental data into a centralized data warehouse. The system enables live vessel tracking, historical analysis, and predictive modeling, while providing tailored access to different stakeholders, including port authorities, shipping companies, yacht owners, and environmental organizations.

Table 1: Overview of Data Sources and Key Technologies

Name	Description	Type	Update
AIS Stream API¹	Provides real-time position reports and event data for vessels within a specified bounding box, including navigational status, heading, and identification information. Ideal for live vessel tracking and maritime safety monitoring.	Stream	Continuous
Marinesia API²	Contains detailed information on all registered vessels and ports, including historical vessel locations, vessel characteristics, port positions, and berth availability. Used for historical analysis and infrastructure mapping.	Batch	Once per day
Open Meteo API³	Offers weather and oceanographic data for a given bounding box, including forecasts for temperature, wind, precipitation, sea level, tides, and wave conditions. Supports operational planning and environmental monitoring.	Batch	Hourly
Apache Airflow⁴	A workflow orchestration tool used to schedule, manage, and monitor data pipelines. Ensures reliable and automated execution of tasks with rich logging and alerting capabilities.	Orchestrator & Batch Data Ingestion Layer	-
Apache Kafka⁵	A distributed streaming platform used for real-time data ingestion, buffering, and delivery to downstream systems, including the data warehouse.	Streaming Data Ingestion Layer	-
Apache HBase⁶	A scalable, distributed NoSQL database designed to store and manage large volumes of structured data for analytics. Used as part of the data warehouse for historical storage and fast retrieval.	Data Storage Layer	-
Apache Spark⁷	A distributed computing framework designed for large-scale data processing and analytics. Supports batch and stream processing, in-memory computation, and machine learning, enabling fast and scalable analysis of maritime and environmental data.	Data Processing Layer	-

¹ <https://aisstream.io/documentation#WebSocket-Messages>

² <https://docs.marinesia.com/features/>

³ <https://open-meteo.com/>

⁴ <https://airflow.apache.org>

⁵ <https://kafka.apache.org>

⁶ <https://hbase.apache.org>

⁷ <https://spark.apache.org>

2 Project Aim

The primary objective of this project is to design and implement a comprehensive maritime data pipeline that enables real-time monitoring, historical analysis, and predictive modeling of vessel movements and environmental conditions. The system integrates multiple data sources, including real-time AIS streams, historical vessel and port information, and marine weather forecasts, into a centralized data warehouse for analysis and visualization.

The specific goals of the project are as follows:

- **Real-time vessel tracking:** Collect and process position reports and critical events from vessels within defined geographic areas, allowing for live monitoring of maritime traffic and safety incidents.
- **Historical analysis:** Maintain a structured repository of historical vessel positions and port visits to enable trend analysis, route optimization, and operational planning.
- **Environmental monitoring:** Integrate weather and oceanographic data, including tides, waves, and atmospheric conditions, to assess the impact of environmental factors on vessel operations and maritime logistics.
- **Data centralization and accessibility:** Ensure that all collected data is harmonized and stored in a scalable, queryable format to support analytics, reporting, and integration with downstream applications.
- **Reliability and fault tolerance:** Implement robust mechanisms to handle interruptions in real-time streams, ensuring continuity of data ingestion and minimizing information loss.

3 Stakeholders

The following stakeholders are relevant for the project:

- **Shipping Companies and Private Vessel Owners:** Access to selected real-time and historical data for vessels they operate or have interest in. This includes position tracking, estimated arrival times, and route history, allowing efficient planning and operational management without exposing sensitive information of other vessels.
- **Yacht Owners and Recreational Mariners:** Receive limited, high-level maritime and weather data relevant to their route and location, such as weather forecasts, tide information, and safe navigation alerts, without full access to commercial or regulatory data.
- **Environmental Monitoring Organizations:** Can access aggregated weather and oceanographic data, including tides, waves, and sea level trends, to evaluate environmental impacts of maritime activity without needing individual vessel details.
- **Data Analysts and Researchers:** Work with anonymized or aggregated historical datasets to study traffic patterns, maritime trends, and environmental correlations, enabling research and predictive modeling without compromising vessel confidentiality.
- **Logistics and Supply Chain Managers:** Access operational data for the vessels and ports relevant to their cargo and shipping routes, allowing optimization of schedules and resources while respecting privacy and commercial restrictions.

4 Data Sources

This section presents the data sources used throughout the project.

4.1 AIS Stream API

This API provides real-time data from the Automatic Identification System (AIS)¹. For a subscribed *Bounding Box*, the following information is available:

- **Position reports** of each vessel, including Latitude, Longitude, Timestamp, True Heading, MMSI², and Ship Name. These reports allow real-time tracking of vessel movements.
- **Ship static data**, including name, call sign, length, width, type, owner/operator, place of build, gross tonnage, destination, and ETA. This provides context for vessel operations.
- **Critical events**, such as Safety Broadcast Messages or Search and Rescue Aircraft Reports, essential for maritime safety monitoring.

Data is accessed via a WebSocket stream. If no messages appear within 3 seconds, the API closes the stream and a new task is initiated to reconnect. All stream data is inserted into a Kafka topic, serving as the primary source for the warehouse and enabling further analysis and integration.

4.2 Marinesia API

The Marinesia API provides detailed information on all vessels registered in AIS. For our project, we retrieve:

- **Vessel information** — given a vessel's MMSI, we can obtain its image, historical location data, and ownership details.
- **Port information** — including location, country, available berths, and images, useful for understanding port infrastructure and docking patterns.

Data is downloaded once a day at 00:00 UTC, mainly for historical vessel positions and an up-to-date port list, which can be integrated with real-time streams.

4.3 Open Meteo API

The Open Meteo API provides weather and oceanographic data for a bounding box. For our use case, the following types of data are retrieved:

- **Sea Levels and Datums** — current sea level relative to reference datums, critical for tides and safe navigation.
- **Weather Stations** — temperature, wind, pressure, and precipitation from nearby stations for real-time monitoring.
- **Astronomical Tide** — predicted tide levels, important for port operations and vessel scheduling.
- **Meteorological Forecast** — short- and long-term forecasts including temperature, wind, precipitation, and storm warnings.

Data is updated hourly, providing a comprehensive view of marine and coastal conditions. This information is integrated into the warehouse to correlate vessel movements with environmental factors.

¹https://en.wikipedia.org/wiki/Automatic_identification_system

²https://en.wikipedia.org/wiki/Maritime_Mobile_Service_Identity

5 Separation of Work

This section outlines the division of responsibilities among the project authors.

Table 2: Separation of Work Among Project Authors

Author	Responsibilities
Paweł Pozorski	Design and implementation of the data ingestion pipeline, integration of AIS Stream API, configuration of Kafka topics for real-time data, and ensuring reliable stream processing with fault tolerance. Integration of streaming sources into data warehouse. Setup of all containers.
Paweł Florek	Historical data management and integration, retrieval and processing of Marinesia API and Open Meteo API data using Apache Airflow, HBase schema design for long-term storage, and integration of batch sources into the data warehouse.
Hubert Sobociński	Workflow orchestration and system monitoring, setup of Apache Airflow for scheduling and automating pipelines, data validation, design of dashboards and reports for stakeholders, and preliminary exploratory data analysis.

6 System Architecture

Figure 1 illustrates the overall architecture of the maritime data platform. The system is composed of external data sources, a Docker Compose environment, storage and processing layers, and orchestration via Apache Airflow.

- **External Data Sources:** AIS Stream provides real-time vessel positions, while Marinesia API and Open Meteo API supply historical vessel/port data and environmental data, respectively.
- **Data Ingestion:** AIS data is ingested in near real-time into Kafka, and batch ingestion DAGs retrieve data from Marinesia and Open Meteo APIs, based on schedule and parameters - existing data from HBase and schedule time frame. An Airflow DAG ensures Kafka topic data is also ingested into HBase in Parquet format.
- **Storage Layer:** HBase serves as the primary data warehouse for historical and real-time data, with HDFS supporting distributed storage.
- **Processing Layer:** Apache Spark performs large-scale data processing, analytics, and report generation on data stored in HBase.
- **Orchestration and Scheduling:** Apache Airflow coordinates all DAGs, scheduling API ingestions, continuous Kafka-to-HBase ingestion, and triggering Spark jobs for data processing and report generation. It also serves as the monitoring interface for the entire pipeline, with alerting for failures.
- **Interactions:** Data flows from external sources to ingestion pipelines, then into storage. Processing jobs access storage for analytics, and orchestration ensures timely and reliable execution of all tasks. Dashed lines in the diagram indicate DAGs that query HBase to determine what data to ingest.

The chosen tools were selected to address the specific requirements of real-time maritime data ingestion, historical storage, and large-scale processing. Apache Kafka enables reliable streaming of AIS data, ensuring near real-time updates. Apache HBase provides scalable, low-latency storage for both streaming and batch data, while HDFS supports distributed storage for fault tolerance and scalability. Apache Spark is used for efficient large-scale data processing, analytics, and report generation. Apache Airflow orchestrates and schedules all ingestion and processing pipelines, ensuring reliable, automated, and maintainable workflows, and also provides notification mechanisms to alert stakeholders in case of failures or important events.

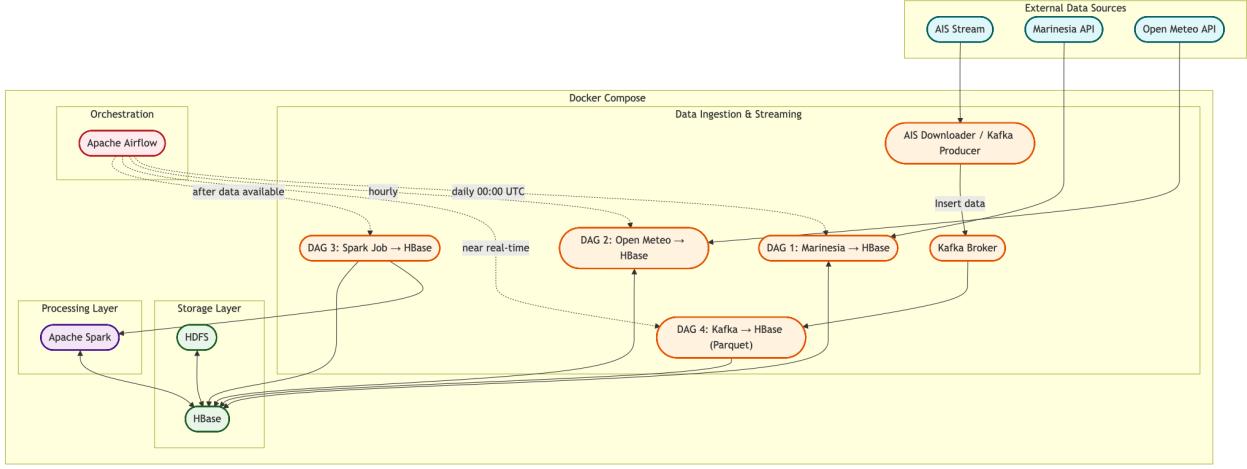


Figure 1: Architecture of the maritime data pipeline with Docker Compose, data ingestion, processing, and orchestration layers.

7 Example Analytical Use Cases

Using the combined data from the Open Meteo weather API, AISStream real - time vessel tracking, and the Marinessia vessel/port registry, a variety of valuable analyses can be performed:

- **Route Optimization:** By correlating AIS - streamed vessel positions and headings with wave height, swell, and wind forecasts from Open Meteo, one can identify which sea routes are more energy - efficient under different marine conditions. For instance, if a particular corridor consistently has high waves during certain times, the system could suggest alternates with calmer conditions, reducing fuel use and improving safety.
- **Safety Risk Monitoring:** With real - time AIS data, we can detect when vessels are navigating in potentially dangerous conditions (e.g., high waves, strong currents from the Open Meteo API). The system may automatically issue alerts (via dashboard, emails, or SMS) for ships in vulnerable regions, enabling proactive safety interventions.
- **Tide Impact Assessment:** Using the tide and sea level endpoints from Open Meteo, combined with vessel location data from AISStream (via Marinessia), one can evaluate how tides affect port approaches, docking, or departure times. Analysts can determine if certain vessels regularly face delays or grounding risk due to tidal extremes.
- **Traffic Pattern Analysis:** With historical AIS data from Marinessia (where position history is available) and static vessel metadata (type, size, owner), analysts can reconstruct traffic patterns: e.g., how many cargo vs. passenger vessels use certain sea lanes, or how vessel traffic changes seasonally. This can inform port planning, environmental impact assessments, or regulatory decisions.
- **Port Throughput & Scheduling Optimization:** By combining port location and berth data from Marinessia with expected vessel arrivals (from AIS) and weather/tide forecasts (from Open Meteo), port operators can better predict busy times, optimize berth allocation, and improve scheduling to reduce wait times and increase efficiency.
- **Historical Trend Analysis and Forecasting:** With historical AIS tracks (from Marinessia) and long-term weather/tide time series (from Open Meteo), one can perform trend analysis: e.g., have wave heights increased over the years in a given sea region? Do certain types of vessels avoid specific areas under particular conditions? These insights can feed predictive models for future maritime risk, traffic, and operational planning.