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

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 API[1]	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 API[2]	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 API[3]	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 Airflow[4]	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 Kafka[5]	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 HBase[6]	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 Spark[7]	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, MMSI[2], 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.

[1] `https://en.wikipedia.org/wiki/Automatic_identification_system`
[2] `https://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 most of the 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 reports in spark 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.

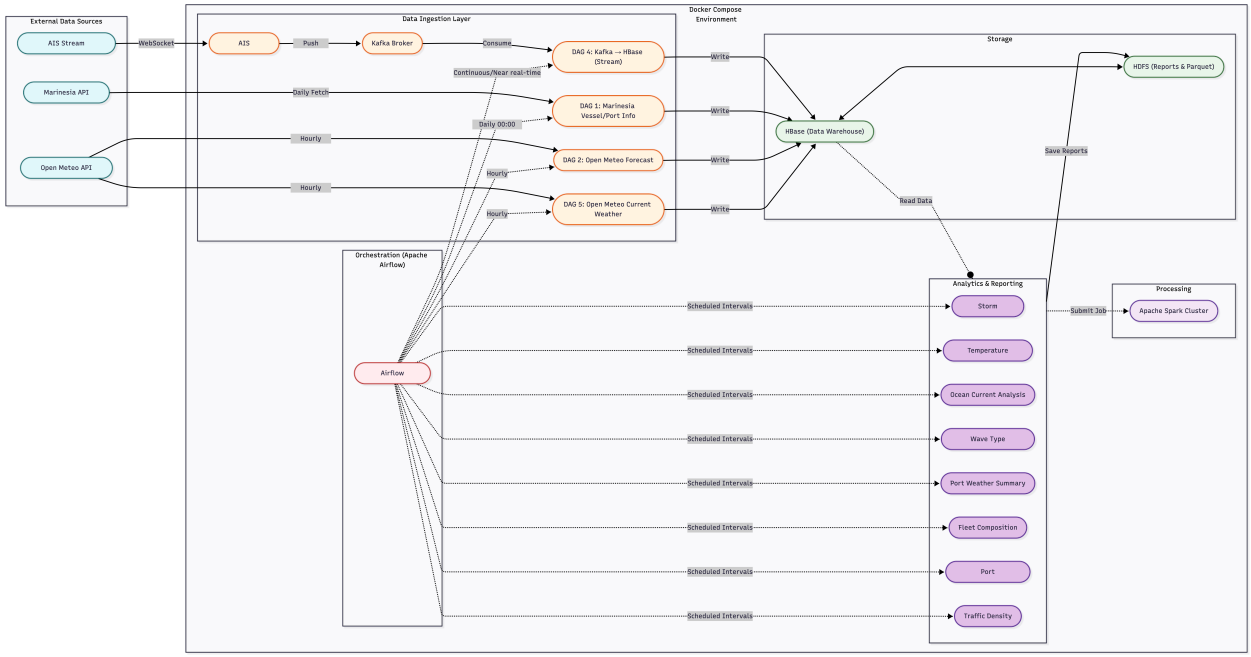

Figure 1: Architecture of the maritime data pipeline with Docker Compose, data ingestion, processing, and orchestration layers.

7 Implementation

This section details the implementation of data ingestion, storage, and analysis processes. Table 3 summarizes the key processes, their data sources, tools used, and scheduling, providing an overview of the system's operational workflow. Details are provided in Sections AIS Stream API, Marinesia API, Open-Meteo API, and Analysis. Analysis is limited to spark views only, without dashboards - they are not part of the project.

Table 3: Schedule of Data Ingestion and Analysis Processes

Process	**Source / Method**	**Tool / DAG**	**Schedule**
AIS Stream Data Ingestion	AIS Stream API (WebSocket)	Airflow DAG, HBase on HDFS	Real-time (continuous)
Vessel and Port Data Fetching	Marinesia API (REST)	Airflow DAG, HBase on HDFS	Daily at 00:00 UTC
Weather and Oceanographic Forecast	Open-Meteo API (REST)	Airflow DAG, HBase on HDFS	Hourly
Current Weather Conditions	Open-Meteo API (REST)	Airflow DAG, HBase on HDFS	Hourly
Storm Detection	HBase (Spark)	Spark job, Airflow DAG	Scheduled intervals
Temperature Analysis	HBase (Spark)	Spark job, Airflow DAG	Scheduled intervals
Ocean Current Analysis	HBase (Spark)	Spark job, Airflow DAG	Scheduled intervals
Wave Type Comparison	HBase (Spark)	Spark job, Airflow DAG	Scheduled intervals
Port Weather Summary	HBase (Spark)	Spark job, Airflow DAG	Scheduled intervals
Fleet Composition Analysis	HBase (Spark)	Spark job, Airflow DAG	Scheduled intervals
Port Characteristics Analysis	HBase (Spark)	Spark job, Airflow DAG	Scheduled intervals
Traffic Density Analysis	HBase (Spark)	Spark job, Airflow DAG	Scheduled intervals

7.1 AIS Stream API

Data from the AIS Stream API is ingested in real-time using a WebSocket connection. Before connecting, the system subscribes to a specified bounding box to receive position reports with entered API key and critical events for vessels within that area. The WebSocket client continuously listens for incoming messages, parsing each message to extract relevant fields such as MMSI, latitude, longitude, timestamp, heading, and ship name etc. Special DAG in Apache Airflow is responsible for processing the incoming stream data. Each parsed message is then saved to specified HBase table on HDFS.

7.2 Marinesia API

The Marinesia API is accessed via RESTful HTTP requests to retrieve vessel and port information. An Apache Airflow DAG is scheduled to run once daily at 00:00 UTC. The DAG consists of tasks to fetch vessel and port data based on predefined bounding boxes. The retrieved data is parsed and transformed into a structured format suitable for storage in HBase. The processed data is then inserted into designated HBase tables stored on HDFS for historical vessel and port information.

7.3 Open-Meteo API

The Open-Meteo API is accessed through RESTful HTTP requests to obtain weather and oceanographic data for saved ports accessed from HBase table stored on HDFS. An Apache Airflow DAG is scheduled to run hourly to fetch the latest weather forecasts, sea levels, tides, and meteorological data. The retrieved data is parsed and transformed into a structured format compatible with HBase storage on HDFS. The processed weather data is then inserted into designated HBase tables for environmental monitoring. Other DAG task is responsible for fetching current weather conditions around the ports and stored with the same way as forecast data.

Figure 2: HBase tables structure for storing ingested data from various sources.

7.4 Analysis

Analytical tasks are implemented using Apache Spark, which processes data stored in HBase on HDFS. Spark jobs are scheduled via Apache Airflow to run at defined intervals. The examples of analysis includes:

- **Storm Detection:** Identifying ports experiencing high wave heights and swell conditions based on predefined thresholds.
- **Temperature Analysis:** Calculating average, minimum, maximum, and standard deviation of sea surface temperatures for each port.
- **Ocean Current Analysis:** Evaluating average and maximum ocean current velocities, along with their variability.
- **Wave Type Comparison:** Comparing wind-generated waves versus swell waves to understand dominant wave.
- **Port Weather Summary:** Generating comprehensive weather summaries for each port, including temperature, wind speed, wave height, and tide levels.
- **Fleet Composition Analysis:** Analyzing the distribution of vessel types (cargo, tanker, passenger, etc.) within the monitored area.
- **Port Characteristics Analysis:** Examining port attributes such as berths, capacity, and geographic location to assess operational capabilities.
- **Traffic Density Analysis:** Calculating vessel traffic density in specified areas.

port_id	port_name	latitude	longitude	avg_wave_height	max_wave_height	avg_wave_period	avg_wind_wave_height	avg_wind_direction
fdd389c54e39cfcb	Triton Offshore T...	57.066667	0.816667	3.11	5.08	7.24	2.57	138.17
0de1ac141a311698	Yme Sls Offshore ...	57.85	4.383333	3.02	4.32	6.82	2.79	136.96
99056c84262c2fd8	Unionhall	51.566667	-9.133333	2.75	4.78	8.71	1.56	248.65
db060de89c1d6d7d	Ustka	54.583333	16.85	2.33	3.56	6.43	2.0	100.0
386d07b3dbc3f224	Traena	66.5	12.083333	2.21	2.66	7.0	1.51	154.56
dde7bec3360ab641	Tyne	54.983333	-1.533333	2.17	3.26	7.64	0.58	119.0
df3a5d380d5b443b	Wladyslawowo	54.8	18.416667	2.04	2.7	6.92	1.54	78.11
8660fe5c41c08352	Visby	57.633333	18.283333	2.03	3.0	5.94	1.94	31.03
202214333e1f5724	Wick	58.433333	-3.083333	1.93	3.0	7.24	0.96	163.61
c47921bde83e2a10	Waterford	52.266667	-7.066667	1.88	4.2	7.86	1.25	247.38

Figure 3: Example Spark analysis result for weather data.

ship_type	avg_length	max_length	avg_beam	avg_draft	vessel_count
Cargo	85.21	338.0	8.99	8.97	135
Tug	13.04	117.0	5.25	5.61	89
Other Type	23.94	90.0	8.02	7.79	62
Tanker	97.08	229.0	9.44	8.97	59
Unknown	45.98	135.0	4.71	4.73	55
Fishing	21.45	38.0	3.76	4.17	42
Passenger	20.9	56.0	6.85	8.46	39
Uncategorized	8.0	36.0	2.32	2.11	37
Pilot	14.05	33.0	6.52	5.43	21

Figure 4: Example Spark analysis result for fleet data.

country	port_count
DEU	347
NOR	231
GBR	219
NLD	106
SWE	94
DNK	85
FRA	75
ISL	38
IRL	34
FIN	26
BEL	26
EST	14
POL	13
HUN	12
HRV	9

Figure 5: Example Spark analysis result for port data.

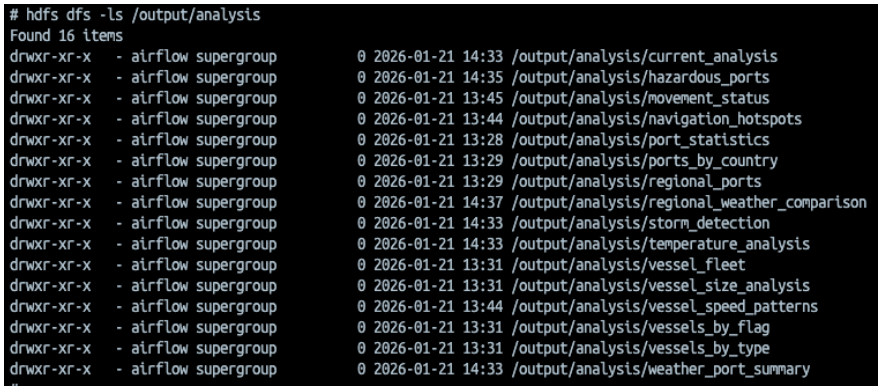

Figure 6: Saved reports in HDFS after Spark analysis.

All reports are later saved back to HDFS for further use and visualization.

8 Future Work Examples of Analytical Use Cases

Using the combined data from the Open Meteo weather API, AISStream real - time vessel tracking, and the Marinersia 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 Marinesia), 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 Marinesia (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 Marinesia 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 Marinesia) 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.