Canalytics: A framework for maritime traffic forecasting

Supervisor: Nasonov D.A., Tereshkin S.E., Butakov N.A.

Team: Canalytics (Bavshin T., Ahmed U., Pyankova M., Lvov D.)

St. Petersburg, 2025

Problem statement

Object - The study of maritime traffic patterns by integrating heterogeneous data sources — such as AIS ship tracking data and maritime-related news — using scalable big data technologies.

Subject - The application of streaming data pipelines, web scraping, and cloud storage to enable continuous ingestion of AIS reports and relevant news, supporting predictive analysis of traffic flow in global shipping chokepoints.

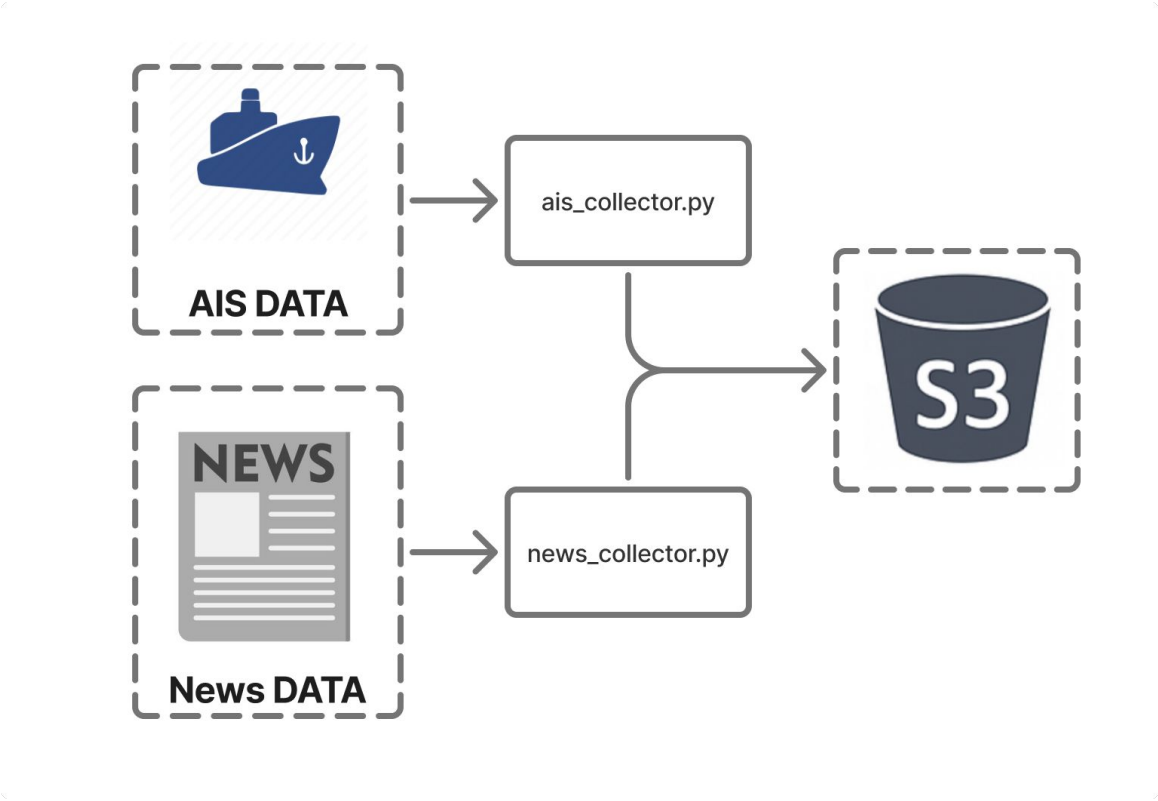
Objective - To develop a data-driven framework for forecasting maritime traffic by building a robust pipeline for collecting and preprocessing real-time tracking and contextual data, and applying analytical methods to identify patterns and predict future traffic dynamics.

Tasks

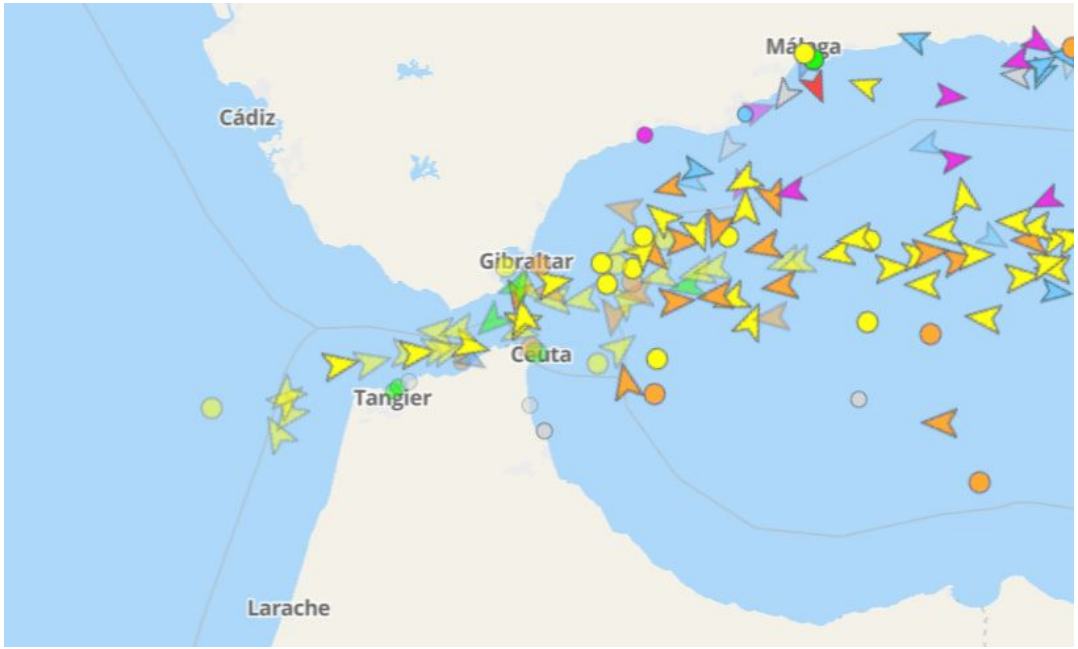
ИТМО

1. Develop a real-time AIS collector using the AISstream WebSocket API
2. Implement a news headline scraper using NewsAPI with maritime keywords
3. Store collected data as raw JSON files in an S3-compatible object store
4. Process data and store in ClickHouse database
5. Orchestrate pipeline components using Apache Airflow and Docker
6. Structure and log data for later ingestion into analytics platforms
7. Perform analysis of the data and present final results

Data collection system



AIS data




```
"Message": {
  "PositionReport": {
    "Cog": 353.2,
    "CommunicationState": 99931,
    "Latitude": 1.2642716666666667,
    "Longitude": 103.713555,
    "MessageID": 1,
    "NavigationalStatus": 5,
    "PositionAccuracy": false,
    "Raim": false,
    "RateOfTurn": -128,
    "RepeatIndicator": 0,
    "Sog": 0.9,
    "Spare": 0,
    "SpecialManoeuvreIndicator": 0,
    "Timestamp": 44,
    "TrueHeading": 511,
    "UserID": 566272000,
    "Valid": true
```


API: https://aisstream.io

News data

HOME // CHINA

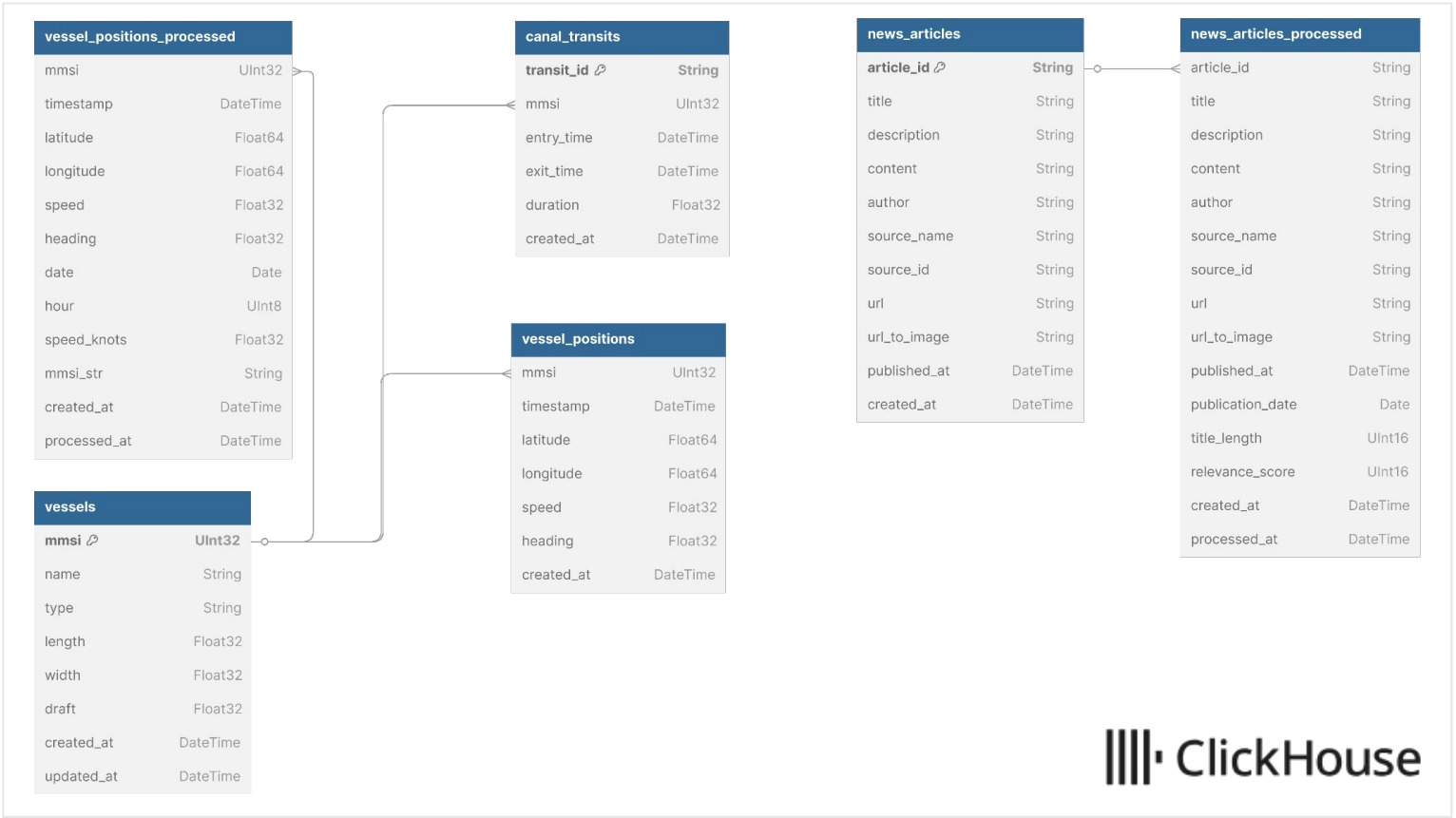
Panama canal crisis: U.S. and China vie for control as Panamanians plead for American support

05/12/2025 // Finn Heartley // 1.3K Views

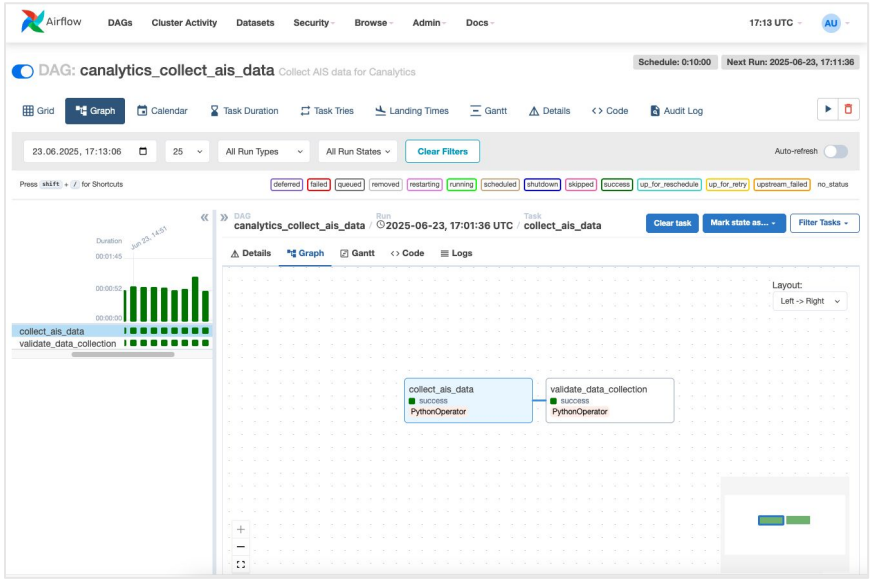

```
"source": {
  "id": null,
  "name": "Naturalnews.com"
},
"author": "Finn Heartley",
"title": "Panama canal crisis: U
"description": "Geopolitical Bat
"url": "https://www.naturalnews.
"urlToImage": "https://www.natur
"publishedAt": "2025-05-12T06:0(
"content": "<ul><li>Geopolitical
```


API: https://newsapi.org

Database structure



Pipeline orchestration with Airflow



We used **Apache Airflow** to orchestrate and automate our end-to-end data pipeline in a modular and maintainable way.

Key Features of Our Orchestration:

- Modular DAGs for:
  - **ais_data_collect**: Real-time AIS stream ingestion
  - **news_data_collect**: Daily scrape of maritime news
  - **etl_process**: Transform and clean raw data
  - **sync_clickhouse**: Load structured data into ClickHouse
- Retry logic and failure handling to ensure fault tolerance
- Task-level logging and alerting
- Dynamic scheduling using DAG intervals
- Fully containerized using Docker & Airflow Web UI

DAG scheduling & intervals

Our pipeline is designed for semi-real-time ingestion and daily aggregation using scheduled DAGs in Airflow.

Why These Intervals?

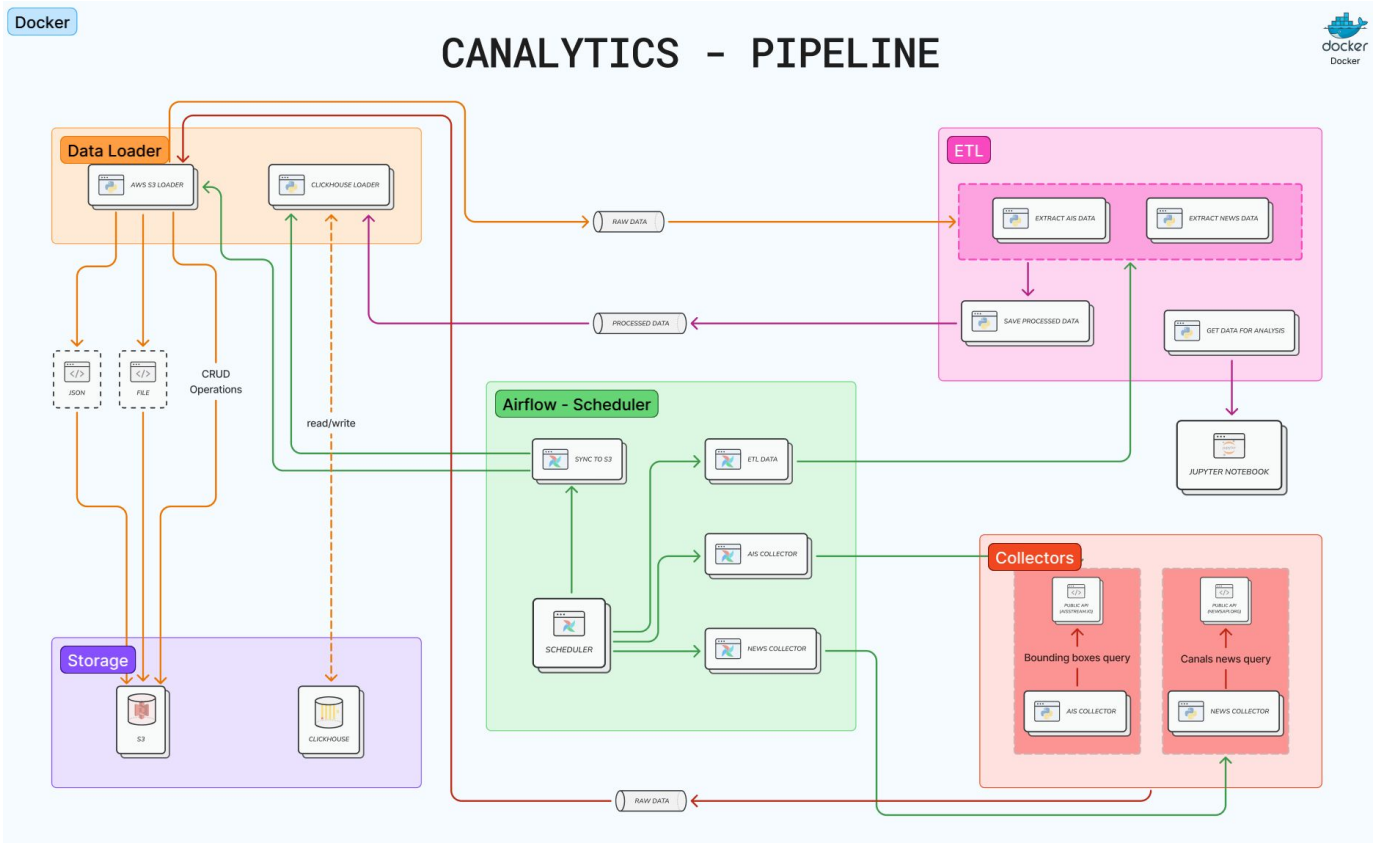
- AIS data changes frequently → frequent polling (10 mins)
- News updates slower → daily ingestion
- ETL and loading jobs → run more often for timely analytics

Features:

- All tasks run automatically based on defined intervals
- Dependencies ensure correct execution order
- Manual triggering available for ad hoc runs

DAG Name	Purpose	Schedule
AIS Collect Data	Collect AIS stream data	every 10 minutes
News Collect Data	Fetch latest news articles	daily
ETL Process	Clean & enrich collected data	daily
Sync Clickhouse	Load data to ClickHouse DB	daily

Project pipeline



Server configuration

VM:
CPU: 4,
RAM: 8Gb
SSD: 30Gb

HyperVisor:
CPU: Intel(R) Xeon(R) CPU
RAM: 256G
SSD: 2Tb

Vessel positions

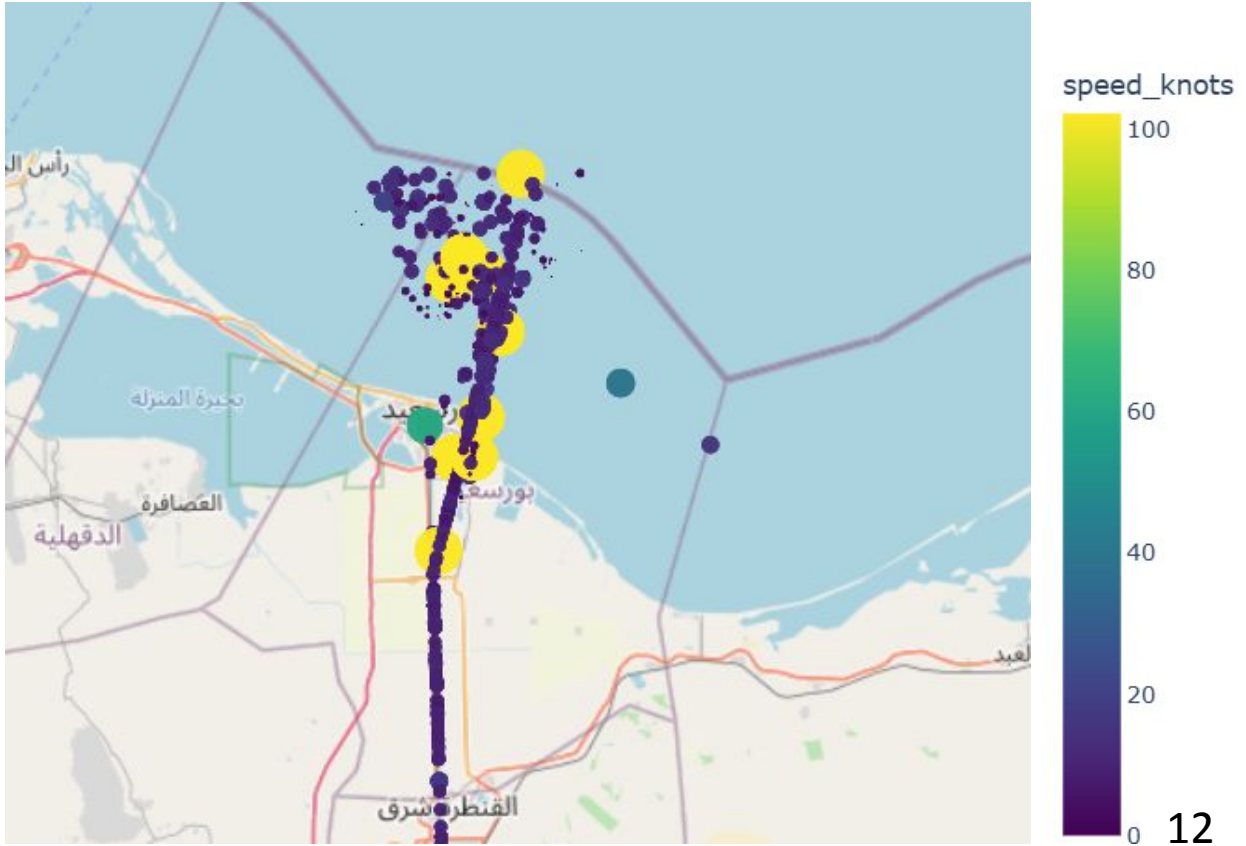
Vessel data shape: (3096294, 12)

- Vessel Data Coverage:

Date range: 2025-05-29 to 2025-06-22
Unique vessels: 5702

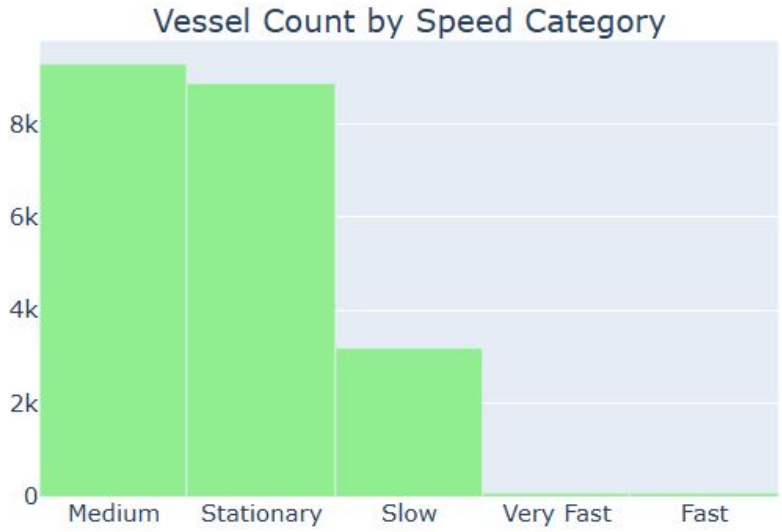
- Suez Canal related:

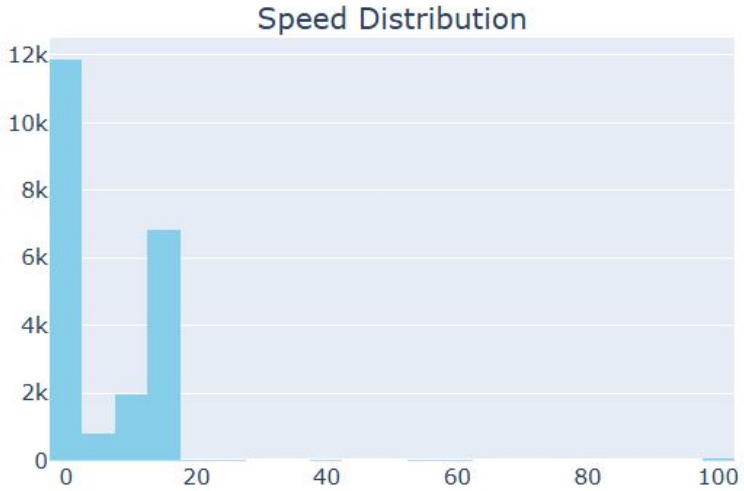
Total positions in area: 21476
Unique vessels in area: 262
Percentage of total data: 0.7%



Speed analysis

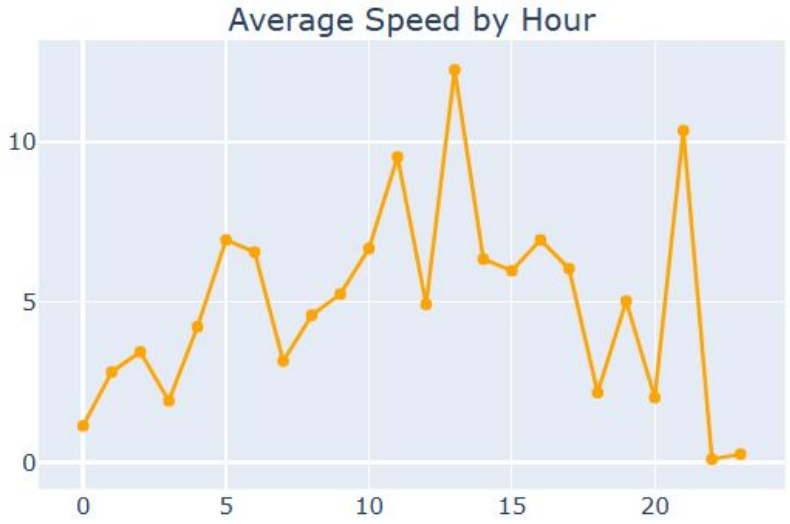
- Average speed: 5.7 knots
- Median speed: 0.1 knots
- Max speed recorded: 102.3 knots
- Vessels stationary (0 knots): 8861
- High-speed vessels (>20 knots): 74

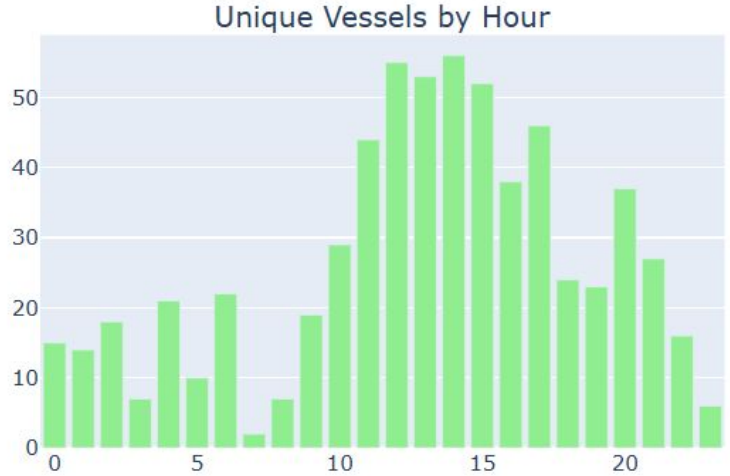




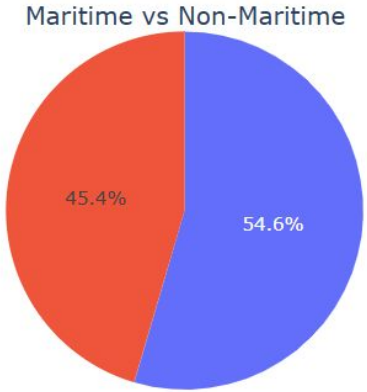
Suez canal traffic

- Peak traffic hour: 12:00 (13937 positions)
- Quietest hour: 07:00 (7 positions)
- Average speed during peak: 4.9 knots





News analysis



News data shape: (253927, 15)

- News Data Coverage:

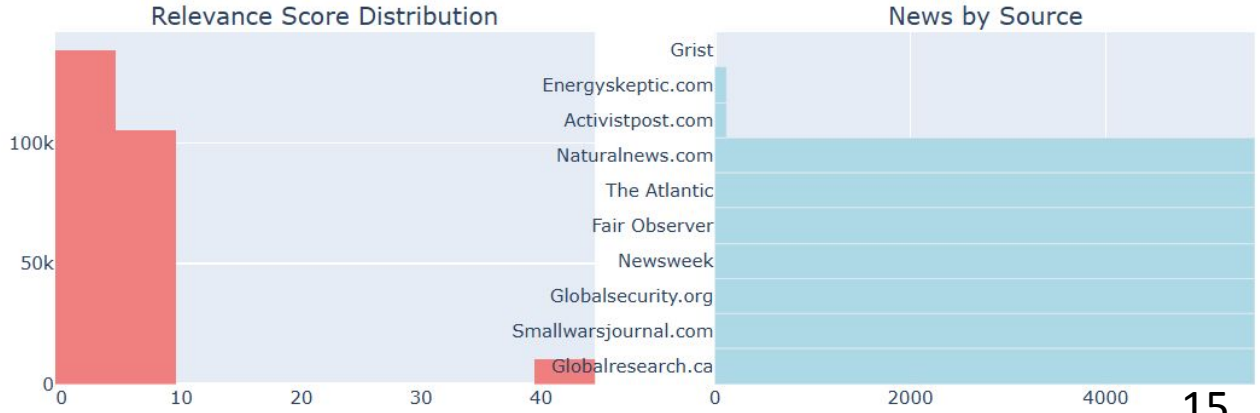
Date range: 2025-05-12 to 2025-06-19
Maritime relevant: 115349/253927 articles

- Suez Canal Related News:

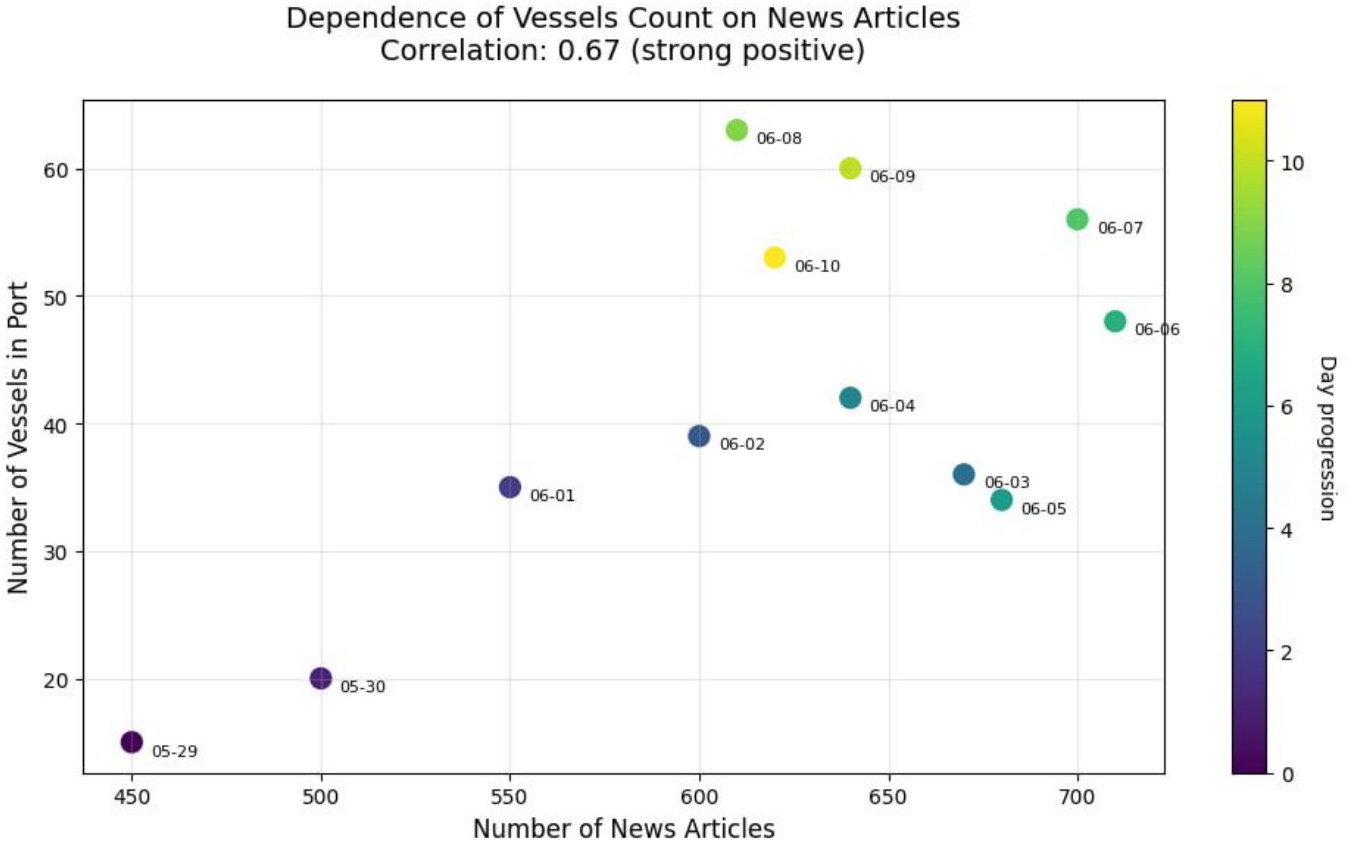
Suez-related articles: 10535
Percentage of total news: 4.1%
Average relevance score: 40.0

<u>Top Suez Canal Related Headlines:</u>

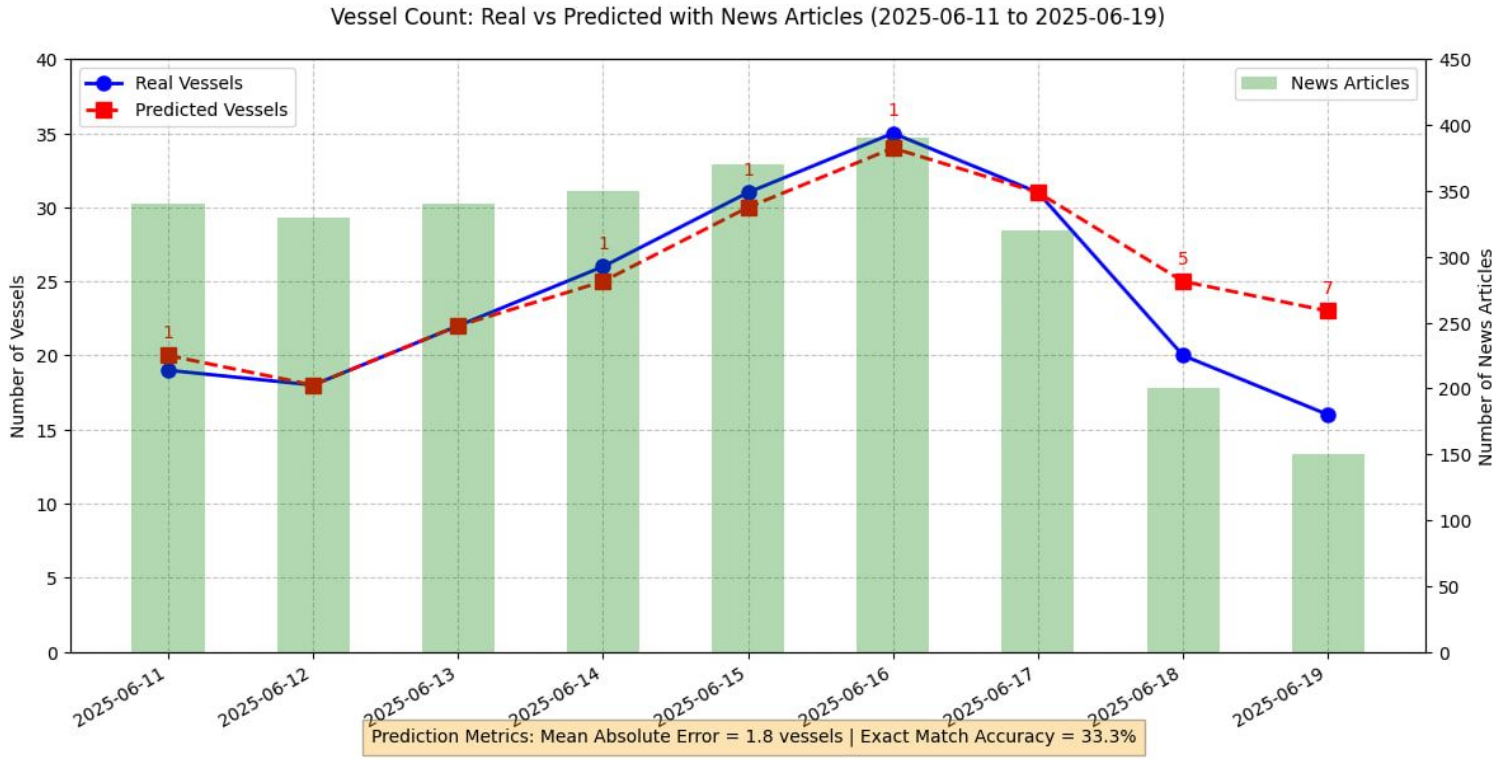
- Panama canal crisis: U.S. and China vie for control as Panamanians plead for Ame...
- Score: 40
- Source: [Naturalnews.com]



Dependence of vessels on articles



Traffic prediction



Results & future work

Data Processed:

- Total vessel positions: 3,096,294
- Suez Canal area positions: 21,476
- Unique vessels tracked: 5,702
- Total news articles: 253,927
- Maritime relevant: 115,349
- Suez Canal related: 10,535

Technologies used:

- S3
- Docker
- Airflow
- Clickhouse
- Jupyter

Future work may include:

- Real-time monitoring dashboards
- Machine learning workflows
- Business intelligence integration
- Automated reporting pipelines
- New data sources (satellite images, stock market fluctuations, etc.)

Thank you!

ITMO more than a UNIVERSITY