The SmartShip project has received funding from the European Union's Horizon 2020 research and Innovation programme under the Marie Skłodowska-Curie grant agreement No 823916

Project Acronym:	**SmartShip**
Project Full Title:	**A data analytics, decision support and circular economy – based multi-layer optimization platform towards a holistic energy efficiency, fuel consumption and emissions management of vessels**
Project Duration:	**60 months (01/04/2019 – 31/03/2024)**

DELIVERABLE 4.1 (final version):

IoT applied tools and technologies and data analytics module

Work Package	**WP4 – Smartship Baseline framework: IoT and advanced data analytics**
Task	**T4.3: Design and Development of Advanced Data analytics module**
Document Status:	**"final v1.1"**
Due Date:	**31/03/2023** (M48 Final version)
Submission Date:	**13/04/2023**
Lead Beneficiary:	**Information Technology for Market Leadership (ITML)**

Dissemination Level	
Confidential, only for members of the Consortium (including the Commission Services)	X

Authors List

Leading Author				
First Name		**Last Name**	**Beneficiary**	**Contact e-mail**
George		Bravos	ITML	gebravos@itml.gr
Anna Maria		Anaxagorou	ITML	aanaxagorou@itml.gr
Co-Author(s)				
#	**First Name**	**Last Name**	**Beneficiary**	**Contact e-mail**
1	Dimitris	Kaklis	DANAOS	dk.drc@danaos.com
2	Ioannis	Kontopoulos	HUA	kontopoulos@hua.gr
3	Vlatka	Katusic Cuentas	CERC	v.katusiccuentas@pontsbschool.com
4	Antonios	Makris	HUA	amakris@hua.gr
5	Konstantinos	Tserpes	HUA	tserpes@hua.gr
6	Dimitris	Panos	BLS	dimitrios.panos@bluesoft.com
7	Jakub	Rola	BLS	jakub.rola@bluesoft.com
8	Dimitris	Bakogiannis	EPSILON	maritime@epsilon.gr
9	Markos	Bonazountas	EPSILON	bonazountas@epsilon.gr

Reviewers List

Reviewers			
First Name	**Last Name**	**Beneficiary**	**Contact e-mail**
Vassilis	Prevelakis	TUBS	prevelakis@ida.ing.tu-bs.de
Ioannis	Kontopoulos	HUA	kontopoulos@hua.gr

Legal Disclaimer

The SmartShip project has received funding from the European Union's Horizon 2020 research and Innovation programme under the Marie Skłodowska-Curie grant agreement No 823916. The sole responsibility for the content of this publication lies with the authors. It does not necessarily reflect the opinion of the funding agencies or the European Commission (EC). Funding Agencies or the EC are not responsible for any use that may be made of the information contained therein.

1. Executive Summary

This deliverable is in line with Article 19.1 of the Grant Agreement 823916 and provides the "IoT applied tools and technologies and data analytics module" of the SmartShip project funded by the Horizon 2020-MSCA-RISE-2018 Action.

The principal aim of SMARTSHIP is to foster knowledge exchange between experts of complementary technology fields (IoT, Data Analytics, Visualization Tools, Optimization Algorithms) applied in the frameworks of Energy Efficiency & Emissions management, towards a holistic framework for energy efficiency and emissions control, thus materializing the next-generation paradigm for the maritime industry. In this context, by capitalizing on available COTS technologies and limited RTD, SMARTSHIP's overall objective is to deliver an ICT & IoT-enabled holistic cloud-based maritime performance & monitoring system, for the entire lifecycle of a ship, aimed to optimize energy efficiency, emissions reduction, fuel consumption, and at the same time include circular economy concepts in the maritime field.

Work Package 4 outputs the SmartShip Baseline framework: IoT and advanced data analytics, lead by ITML (M10-M60). This deliverable is based on the output of T4.1 and T4.2 as well as the architecture from WP3 and the requirements of WP2. ITML will design and develop the IoT-based data analytics module of SmartShip, which will be the core of the multi-level optimization of the vessels' operation and management in terms of fuel consumption, energy efficiency, emissions, and circular economy principles.

This deliverable was developed by the SMARTSHIP Project in order to document the procedures from the demonstration which concerns the "IoT applied tools and technologies and data analytics module".

The document contains the structure and procedures regarding the specific deliverable, and initially provides an overview of the project baseline framework along with its objectives and the relation to the work programme.

List of Contents

List of Figures

List of Tables

List of Acronyms and Abbreviations

Term	Description
GA	Grant Agreement
WP	Work Package
PO	Project Officer
KPI	Key Performance Indicator
IoT	Internet of Things
ICT	Information and Communication Technology
ADM	Administrative
ESR	Early-Stage Researcher
ER	Experienced Researcher
DoA	Description of Action
GPS	Global Positioning System
AIS	Automatic Identification System

2. Introduction

2.1 Scope and objectives of the deliverable

The scope of the deliverable is the design and development of the IoT-based data analytics module of SmartShip, which will be the core of the multi-level optimization of the vessels' operation and management in terms of fuel consumption, energy efficiency, emissions and circular economy principles. Furthermore, in this deliverable it's important to identify and report of any market-ready tool and technology already applied in the maritime industry, related to IoT-based advanced data analytics.

2.2 Structure of the deliverable

The document contains the structure and procedures regarding the specific deliverable, and initially provides an overview of the project baseline framework along with its objectives and the relation to the work programme.

The deliverable consists of seven main sections:

First section includes the executive summary of this document.

Introduction part of this deliverable, stating the scope and the objectives is addressed in the **Second section.**

Third section focuses on the state of the art in Advanced data analytics and IoT Technologies. This section helps the viewer to define the concept of Internet of Things, along with the Advance Data Analytics theory.

Fourth section analyzes the conceptualization of IoT technologies and state of art in Advanced data analytics in the Maritime industry.

Fifth section describing the design and development of Advanced Data analytics module, which includes three sub sections along with figures and modelling snapshots.

Finally, the Conclusion part of this deliverable appears in the **Sixth section**. References used in this deliverable are listed in the **Seventh section**.

2.3 Relation to Other Tasks and Deliverables

This document references procedures which are described in detail in D3.1 "SmartShip circular economy-based functional architecture", submitted in M36 (March 2022). Subsequently, the parts of Design and Development of Advanced Data analytics is based on the inputs from T4.1 and T4.2, as well as the architecture from WP3 and the requirements from WP2.

3. State of the art in Advanced data analytics and IoT Technologies

3.1 Internet of Things (IoT)

Today's industries face the problem of aggregating and analyzing data consumed from multiple heterogeneous data sources. Such data sources could be any (Internet of Things) IoT device, from Raspberry PIs to sensors on board a moving object. In the maritime domain, things are no different as most of the vessels carry AIS transponders and GPS devices. Furthermore, vessels are also equipped with sensors attached to their engines and other main functionality components. Taking into account the total number of vessels globally, the problem of data aggregation and consumption becomes even harder to tackle.

To this end, several works have focused on the domain of data streams aggregation. In order to tackle problems posed with centralized approaches, the authors in [1], present a novel resource-aware network-partitioning algorithm which is able to partition and distribute data based on the load of each node over the network and the change of data stream rates. The developed Distributed Stream Management Infrastructure (DSMI) supports an SQL-like query language to describe the process of aggregations for producing a new aggregated data stream from existing data streams. Similar to [1], authors in [2] present a novel stream join model, called join-biclique. Based on this model which treats an entire cluster as a bipartite graph, a distributed stream join system is developed, called BiStream. This system supports efficient full-history joins, window-based joins, online data aggregation and resource management for scaling. On the other hand, Babcock et al. [3] developed algorithms able to determine at what points in a query plan, load-shedding should be performed and the amount of load to be shed. The main idea behind this approach is when the system resources cannot deal with the amount of data being consumed at a given time, system load must be reduced by dropping unnecessary data tuples. Similarly, authors in [4] present storage-efficient algorithms for decay functions which determine the relative contribution of each data to the aggregate. The contribution is highly related to the time passed from the moment the data was generated.

Another approach in stream aggregation and consumption is to treat each IoT device as a microservice. Butzin et al. [5] investigate different patterns and aspects in the microservices approach and examine how these practices can be integrated in the IoT domain. In the microservice architecture, individual distributed interconnected services are designed to work together and structure an application. The interoperability of IoT services and the creation of value-added applications could benefit by employing the same architectural design. The aspects compared, related to self-containment, monitoring and fault handling. Self-containment property focuses on separation of the functionality and enforces isolation via independently deployable units. By adopting this property in IoT, several benefits arise such as independent evolution of services, easier deployment and better decoupling between services. Monitoring is a process of reporting, gathering and storing information. Each service should provide an interface about its health status in order to prevent other services to call a broken one. Microservices and IoT employ the concept of circuit breaker in conjunction with the load balancer pattern. The circuit breaker prevents messages delivered to broken services and enables the load balancer to distribute the workload only on "healthy" services. In conclusion, this research work supports those architectural goals of microservices and IoT are quite similar and IoT could benefit from aspects used in the microservices approach.

Following this direction, [6] presents a vision of applying microservice architecture in an IoT system. Several challenges concerning IoT systems have been already addressed and the Internet consists the backbone for the IoT. However, existing IoT systems facing several well-known problems including interoperability, security flaws, heterogeneity of technologies and protocols used, power limitations etc. In this research, "things" are not treated as atomic elements of the system but rather follow the SOA approach where IoT is a network of services. An IoT node is a smart object that provides services over the network. Thus, the focus is shifted to the level of data and services rather on devices and communication. However, the SOA approach is heavyweight and consists of centralized service models.

As a solution the microservice pattern is applied to IoT systems where each component is independently developed and deployed. Because IoT systems have important differences with cloud-or web-centric patterns, the microservice pattern is combined with complementary patterns which are able to solve several issues concerning the Internet of Things. These patterns include API Gateway, distribution, service discovery, containers and access control. Two case studies are employed, and the results show that in order to successfully apply microservices in IoT systems, many trade-offs should be considered and open questions to be addressed.

3.2 **Advanced Data Analytics**

To analyze and process large volumes of data consumed from multiple sources, researchers' interest has shifted focus towards extracting and discovering knowledge in an automated fashion. With the increase of tracking sources, several studies have focused on surveillance, tracking and route monitoring of moving objects. The research fields in surveillance include classification and clustering of moving objects' trajectories.

Trajectory classification methods extract different features from the spatiotemporal properties of trajectories, such as speed, acceleration and direction change, to use as input for trajectory classifiers. As referred in [7], the difference between the classification techniques is focused on the type of trajectory features extracted for creating the classification model. Effective trajectory classification requires generating a set of features that discriminate against the class. Therefore, the attention in [8] is focused on trajectory feature generation. Two assumptions that are taken into consideration is that discriminative features are more likely to appear as parts of trajectories (sub-trajectories) and as regions. In this work, a feature generation framework called TraClass is presented. The framework generates a hierarchy of features by combining two types of clustering: i) region-based which discovers regions of trajectories that belong to one class and ii) trajectory-based which discovers sub-trajectories that present common moving patterns of each class.

TraClass uses a multi-resolution grid structure, which divides the space into a finite number of cells. The cell sizes are reduced until the trajectories inside the grid belong to the same class. Each trajectory is segmented into a set of partitions which are grouped using a density-based clustering method similar to the DBSCAN algorithm. If the trajectories in a grid cell belong to the same class then it is selected as a feature otherwise the trajectories are split by direction change. The grid cells and sub-trajectory constitute the features of the classification model and belong to either a region-based or a trajectory-based cluster. The combination of two clustering methods achieves better classification results in terms of accuracy. In [9] a classification method is presented, which computes and analyzes features in both spatial and temporal domains. The proposed method consists of three stages and segments the trajectories by using two types of grids. In the first stage, the trajectories are partitioned based on their space location, and the time duration of the sub-trajectories inside each grid is calculated. In the second stage, the trajectories are partitioned based on temporal windows of increasing sizes, and features related to average and standard deviation of speed, acceleration, turning angle and traveled distance are extracted from the sub-trajectories inside each time window. In the third stage, features from the spatial and temporal domain are employed as input for the trajectory classifier. In [10] the Nearest Neighbour Trajectory Classification (NNTC) is used for trajectory data. For data enrichment, an initial pre-processing is performed which adds additional features such as the time of the next position in the sequence, the time interval and the distance in space to the next position, speed, direction, acceleration and direction change. Then, trajectories are represented as the sequence of these features and assigned to the same class as its neighbour (closest trajectory). NNTC predicts the classes based on the label of the nearest trajectory. A segmentation and feature extraction method for trajectories which considers local and global features is presented in [11]. The proposed method calculates features every two consecutive trajectory points and then the trajectories are represented as sequences of each feature. This method strives to classify the movement characteristics of different types of dynamic objects and to extract possible similarities among the movements. Local features are extracted from sub-trajectories with the same characteristics and global features are statistics of the entire trajectory. The vast number

of global and local statistical descriptors which can be used as features in the classification process present correlations. Thus, the principal component analysis (PCA) is used for selecting the best features. A trajectory classification method which is based on shapelet analysis is presented in [12]. The method extracts relevant sub-trajectories called movelets, in order to generate local features of each trajectory and compares the distance of each sub-trajectory to all trajectories in the dataset.

J'unior et al [13] presented an active learning approach called ANALYTiC, which enables semantic annotation on the learning set. The proposed approach computes the speed, the direction variation and the traveled distance between the consecutive points of a trajectory, and it calculates the global features of minimum, maximum and average values of all point features to be used as the classifier's features.

Bolbol in [14] presents a framework for classifying the GPS segments into transportation modes (car, walk, cycle, underground, train and bus). The proposed framework is based on Support Vector Machines (SVMs) classification and extracts features as the average acceleration and average speed of the trajectory. SVMs are easily trained and can be applied directly to the data without any prior feature extraction process in comparison with other machine learning methods. In the first step, an initial process which segments the track is performed and then the data are passed to the SVM learning process. The proposed method segments the trajectories in a pre-defined number of sub-trajectories and a fixed-length moving window is applied to cover a certain number of consecutive segments in order to clarify different transportation modes. A Convolutional Neural Networks (CNN) architecture for trajectory mode classification based only on raw GPS trajectories, is presented in [15]. Initially a pool of attributes/features (e.g speed, acceleration, direction change, and stop rate) are computed from sequential trajectory points. Then, the trajectories are represented by a vector of four dimensions, one for each feature. This vector is fed into the CNN to estimate the transportation mode. In order to evaluate the performance of the proposed approach, the CNN model was compared with traditional machine learning algorithms including K-Nearest Neighborhood (KNN), RBF-based Support Vector Machine (SVM) and Decision Trees (DT). Furthermore, authors in [16] propose a supervised learning approach for transportation mode classification based on user's GPS logs. In addition to simple velocity and acceleration, the method identifies a set of features which are more robust to traffic condition. Specifically, for each trajectory are extracted features such as length, maximum speed and acceleration, average expectation, and variance of speed, heading change rate, stop rate and velocity change rate. Subsequently, these features are used for training a Decision Tree-based model in order to perform predictions.

Finally, authors in [17], fused a Genetic Algorithm (GA) with two other algorithms, General Hidden Markov Models (GHMM) and Structural Hidden Markov Models (SHMM), for the classification of trajectories and evaluated their approach in two different surveillance datasets, MIT car [18] and T15 [19].

In the maritime domain several approaches have been studied regarding the field of trajectory classification. Initially, studies exploited data originated from the Vessel Monitoring System (VMS) to classify fishing activity [20], [21], [22], [23], a satellite-based monitoring system which provides location, course and speed of vessels to fisheries authorities at an one-hour time interval. Walker et al. [20], [22], presented a Bayesian state-space model to classify VMS data of tuna purse-seiners into three different activities, fishing, tracking and cruising. Each state was assumed that it followed an order one Markovian process and the prediction of the activity corresponded to the state that had the maximum a posteriori probability to occur. Authors in [23] proposed a set of features that are representative of trajectories of different fishing vessel types. These features are then used to feed machine learning schemes of XGBoost with the aim of distinguishing between nine different fishing vessel types, achieving a classification accuracy of 95.42%.

However, after the adoption of the AIS from the International Maritime Organization (IMO) as a mandatory means of vessel monitoring which covers a wider range of vessels and has higher transmission frequency compared to the VMS, studies focused more on data collected from AIS receivers. Mazzarella et al. [24], analyzed the behaviour of fishing vessels by detecting the stops and

moves in their trajectory. To this end, they combined two algorithms, namely CB-SMoT [25] and DB-SMoT [26]. The former identifies the speed variations in a trajectory and the latter identifies the change of its direction. As a next step, they identified clusters with the use of the density-based algorithm DBSCAN, each cluster indicating a dense area of fishing activity. Another usage of the DBSCAN algorithm can be seen in [27] where authors detect through clustering Points of Interest (POI) in the vessel trajectories which are then used to extract features that are fed into a classifier.

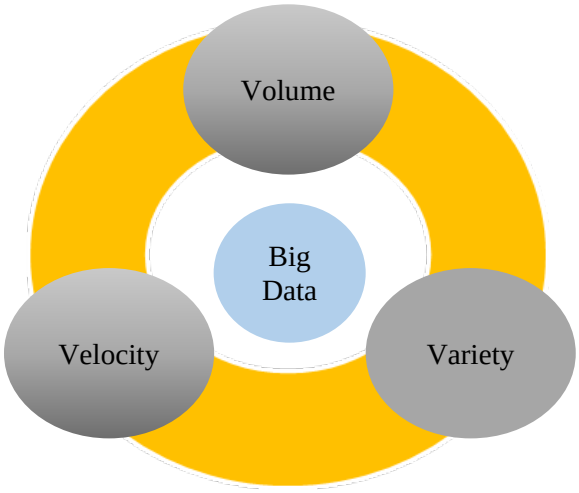
Souza et al. [28] presented three classification techniques to determine whether specific types of fishing vessels are engaged in fishing activity or not. The analysis focused on three types of fishing vessels, each one using a different type of equipment for fishing: i) trawlers, ii) longliners and iii) purse-seiners. Based on their trajectory behaviour when engaged in fishing, authors determined that in each vessel type, a different classifier is suitable and therefore used three different classification models. Each classification model identifies parts in the trajectory that correspond to either fishing activity or not (when the vessel is not fishing, it is either moving towards the fishing area or from the fishing area). The disadvantages of their proposed methodology are that each classifier performs a binary classification task and that the gear type is not always given by the AIS, making it harder to choose a proper classifier in a real-world application setting. Finally, Jiang et al. [29] presented classifiers which use neural networks and autoencoders achieving a high-accuracy classification performance. Their methodology is similar to that of [28], in the sense that they perform binary classification to detect when a specific type of fishing vessel is engaged in fishing activity. Their methodology was evaluated in longline fishing vessels, a fishing type that uses long nets and hooks attached to them to capture fish and compared with other classifiers such as Random Forests and SVMs.

4. IoT based data analytics tools and technologies applied in the Maritime Industry

4.1 Introduction

4.1.1 Big Data in the Maritime industry today

According to the definition of big data[1], it is the name given to the large volume of structured and unstructured data produced in our personal and professional lives. It can be defined by its variety, velocity, and volume with which it is generated. Big data analysis is exceptionally advantageous since it allows businesses to expose hidden patterns, unknown correlations, uncertainties, market trends, and other meaningful information. Big data offers great capabilities to optimize operations to chime with ship calls, renew port assets, and ensure optimum cyber-security.



SEQ Figure: Big Data

Big data has the possibility to transform the Maritime Industry. Through applications and insights, big data is deploying new opportunities to drive innovation and deliver tangible operational efficiencies across the shipping world. However, raw data is not enough, analysis of this data will provide information that will allow the Maritime Industry to move forward. Thus, applying big data offers huge potential in Maritime Industry, and an extensive review follows analyzing the benefits, challenges, and initiatives.

4.2 Current and potential application areas

Nevertheless, organizations leverage diverse data pools to drive value, and big data has significantly benefited industries such as finance, media, telecom, and healthcare, its uptake by the maritime industry has been slow. According to a report by Ericsson[2], the maritime industry lags behind other transport industries in terms of its use of information and communications technology.

There are numerous benefits that the industry can derive from the use of big data. The industry generates roughly 100-120 million data points daily from different sources, such as ports and vessel movements. Companies can analyze these data points to identify efficiencies such as faster routes or ideal ports.

Big data remains untapped in the shipping industry. Therefore, there are huge opportunities for innovation, usage, optimal performance, and better leveraging assets.

The following table presents a snapshot of application areas for big data in the maritime industry:

[1] https://www.investopedia.com/terms/b/big-data.asp

[2] https://www.ericsson.com/en/press-releases/2015/1/maritime-ict-cloud-enables-ships-to-join-the-_networked-society

Table 1. Application areas for Big Data in the Maritime industry

ROLE	FUNCTION	EXAMPLE OF BIG DATA APPLICATION
Ship Operator	Operator	• Energy saving operation • Safe operation • Schedule management
	Fleet planning	• Fleet allocation • Service planning • Chartering
Ship Owner	Technical Management	• Safe operation • Condition monitoring & maintenance • Environmental regulation compliance • Hull and propeller cleaning • Retrofit and modification
	New building	• Design optimization

The following are some key application areas for big data in the maritime industry:

Table 2. Key Application Areas for Big Data in Maritime industry

AREA	DESCRIPTION
Chartering	The most crucial thing for Charterers is to get the proper ship for cargo at the most cost-effective price. Big data analytics could provide charters with available, precise, and useful information to enhance the decision-making process. Charters can incorporate Automatic Identification Information (AIS), estimated time of arrival/departure (ETA/ETD), vessels details, position reports, and market information into an exchange portal to find all available alternatives, as well as freight forecast i.e., BDI.
Operations	Speed is all about fuel consumption. Operating a vessel at its optimum speed is tough as it changes over time due to various factors such as engine and maintenance. Big data analytics could assist ship owners in determining the ideal speed for fuel consumption, considering factors such as bunker expenses, freight rates, and schedules. Fuel consumption information/data can also be provided for cost-benefits analysis of vessel's maintenance.
Voyage Operations	Voyage managers, terminal operators, and port agents must know the estimated time of arrival (ETA) and cargo information. There is the possibility of tracking the vessel using dashboards instead of relying on notes, emails, and phone calls. In this way, there is significant information on practical decisions about terminal and berth allocation, cargo handling, and route tracking.
Vetting	Another important matter is the vessel acceptance by Characters. Ship owners must take care of their ships in order their fleet to pass the selection criteria. Vetting includes receiving feedback from various entities (Port Authorities, terminals, inspectors, etc.). In this case, data analytics could assist charterers, and

	vetting firms analyze the different sources of information, navigation, and safety management.

4.3 The future of big data in the Maritime Industry

The following table presents the key trends in the use of big data in the maritime industry:

Table 3. Key trends in the use of Big Data in Maritime industry

TREND	DESCRIPTION
Technology capabilities are developed through Partnerships	Big data analytics are implemented through collaboration between shipping companies, technology suppliers, institutions, and universities. Collaboration can unlock synergies to generate direct value for customers/end users creating a unique ecosystem.
Through big data could be achieved the bunker cost reduction	One of the most crucial matters in the Maritime industry is fuel consumption/prices. Ship Owners and Operators are trying to eliminate their bunker costs. The use of big data through maritime software, there will be performance monitoring and optimization solution for fuel energy reasons.
Maritime companies are willing to set up internal infrastructure for big data execution	Maritime companies are developing internal platforms and entities to ensure efficiency, forecasting, and data security.
Big data entry in shipping is supported by funding	There are many types of funding (i.e., EU H2020 Calls, tenders, EMFF programs) in order to boost the use of big data in different applications of shipping.

4.4 Challenges for Maritime Industry of Big Data Analytics

According to the "Challenges and Opportunities of Big Data Analytics for Upcoming Regulations and Future Transformation of the Shipping Industry"[3], a research article of the maritime industry, it generates an enormous amount of data from multiple sources and in different formats, including traffic, cargo, weather, and machinery data. The volume and variety of data continue to increase day by day due to the application of sensor technology in the industry. Big data analytics are new to the maritime industry and address many issues, such as adaptability and integration.

There are a lot of challenges that the maritime industry faces in terms of big data; some of them are described below.

[3] Zaman, I., Pazouki, K., Norman, R., Younessi, S. and Coleman, S., 2017. Challenges and opportunities of big data analytics for upcoming regulations and future transformation of the shipping industry. Procedia engineering, 194, pp.537-544. DOI:10.1016/j.proeng.2017.08.182

- **Data Transfer:** Ships typically have a very large number of sensors onboard. A major cause of uncertainty comes from data transfer from those sensors. Every sensor requires a specific communication bandwidth, so it is important to have appropriate data communication for the individual sensor to transmit the information to the database. The data transfer speed may be accelerated with the help of high-tech communication systems.

- **Cybersecurity:** This is a burning issue for any IT system. The data network's safety, security, and management will become vital for future shipping. This will need to be protected from external interventions such as piracy, viruses, or terrorist attacks. Cyber security will be the key issue for any naval system to prevent corruption in maritime security. A cyber-attack on the sensor network would interrupt the overall system and could be responsible for significant losses in the business.

- **Data Quality:** Low-quality data would potentially lead to errors in interpretation. The database will not be able to keep track of all new entries. Therefore, ideally, the data should be error free. Data quality will be a big concern for the industry.

- **Data Integration:** The current data collection systems in the marine industry are inconsistent and often unreliable. Data from different sources will need to be integrated for analysis. For example, fuel consumption, GPS data, and engine data would need to be integrated to monitor the vessel's performance.

- **Data Ownership:** Ownership allows access to the data to read, create, update, and delete database entries and allows traceability through the data lifecycle. The shipping industry is based on a complex supply chain; stakeholders include ship owners, operators, customers, port authorities, and Classification Societies. Ship operators will have access to the full set of machine data, and Classification Societies will get access to data for safety or classification purposes. Port state authorities will require access to cargo and personnel information. Ownership of data is crucial to the shipping industry, and it will become more challenging for ship operators to distribute the data ownership and the level of authority in the future.

- **Data Protection:** Data will move between individual parties because of different interests. Sensitive data will probably need to be shared externally, prioritizing security and privacy for data protection and maintaining data quality.

- **Adoption and Standard Management:** The industry has to look forward to adopting big data analytics to understand the hidden features and benefits of using the available data. The shipping industry will need to create an environment and awareness across the stakeholders to adopt new technologies, tools, and processes and regulate standards.

- **Human factors and Practice:** Increasing the connectivity between the crew and shore staff in shipping companies will become more important. The data transfer between a ship and shore and from shore to ship will increase to drive toward optimal operational efficiency and safety. The ship and shore personnel will be required to undertake additional training to provide support for this.

- **Business Model:** The shipping industry is moving towards significant technological change. This will lead to a change in the business model of the industry. New business models will enable the development of a transparent industry associated with transferring knowledge and data-driven systems.

4.5 Internet of Things Technologies in Maritime Sector

The suite of technologies enabling the Internet of Things promises to turn almost any object into a source of information about that object. Sensor technology allows objects to have the "perception"; RFID technology makes them "speak"; machine-to-machine (M2M) lets them “exchange”; finally, IoT allows all objects in the world to interconnect. IoT solutions thus enable businesses to analyse data generated by sensors on physical objects in a world of intelligent, connected devices. Therefore, IoT connects the digital and physical worlds by collecting, measuring, and analysing data to predict and automate business processes. IoT, can also help accelerate the transition to a circular economy in the industry.

The convergence of IoT, cloud, and big data, create new opportunities for analytics towards a completely new paradigm of big data analytics. This creates a new way to differentiate products and services and a new source of value that can be managed in its own right. Realizing the IoT’s full potential motivates a framework that captures the series and sequence of activities by which organizations create value from information: the Information Value Loop.

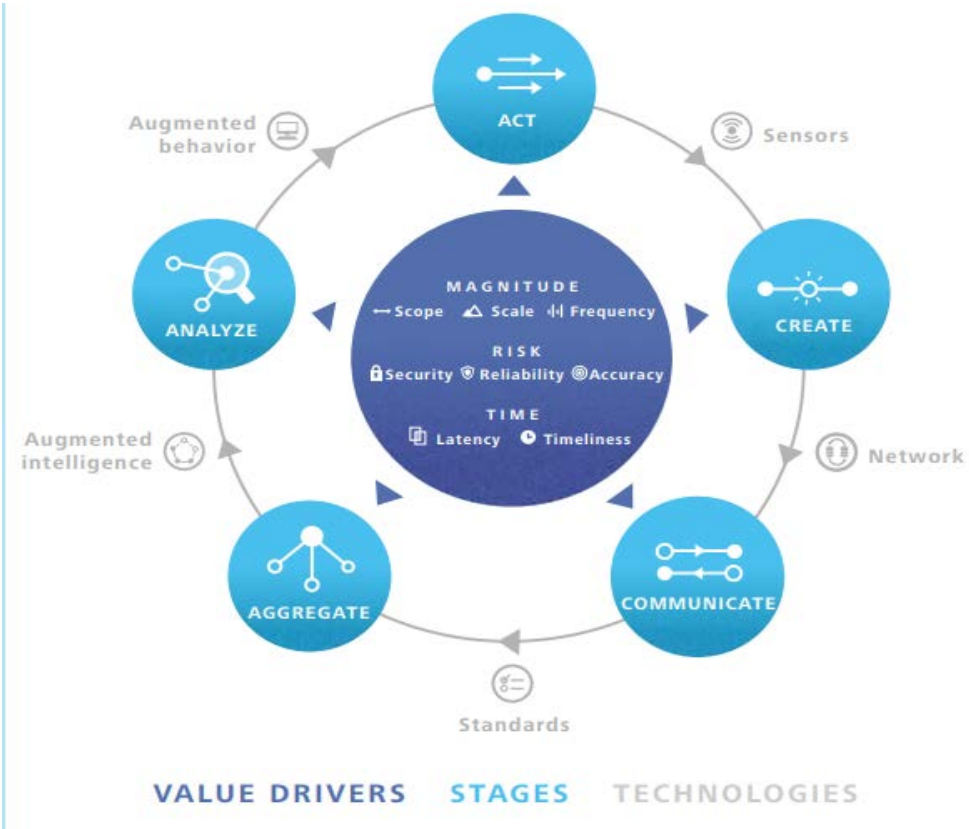


Figure 1. The Value Loop

As the next big leap in mobile and wireless communications, 5G is expected to open numerous possibilities in maritime communication. 5G could potentially optimise the routing undertaken by maritime vessels, resulting in less distance travelled and lowered emissions.

Real-life applications include the introduction of smart drones for real-time monitoring, ship-shore communication for **vessel traffic management,** and **just-in-time operations**. Furthermore, maritime 5G will facilitate the adoption of **autonomous vessels** with low latency connectivity for remote operation and hasten Internet-of-Things sensors during **search-and-rescue** for real-time communications and accurate positioning.

Industrial IoT (IIoT) can realize an immediate return on investment (ROI). Sensors on the ships work together with sensors at the docks to monitor the volume of goods and unloading speeds. The data obtained from the sensors are fed through sophisticated applications which utilize big data analytics to determine a realistic ETA. Intelligent IoT Messaging technology provides the ability for anyone to query the ETA status of the ship.

Cyber-Physical Systems (CPS). These systems are combinations of several major innovations in digital technology poised to transform the industry. The technologies include cloud computing, the Internet of Things (IoT), Blockchain, sophisticated sensors, data capture and analytics, advanced robotics, and artificial intelligence.

The core vision is to enable seamless information exchange to streamline transport operations, increase safety, improve competitiveness, and reduce the environmental impact.

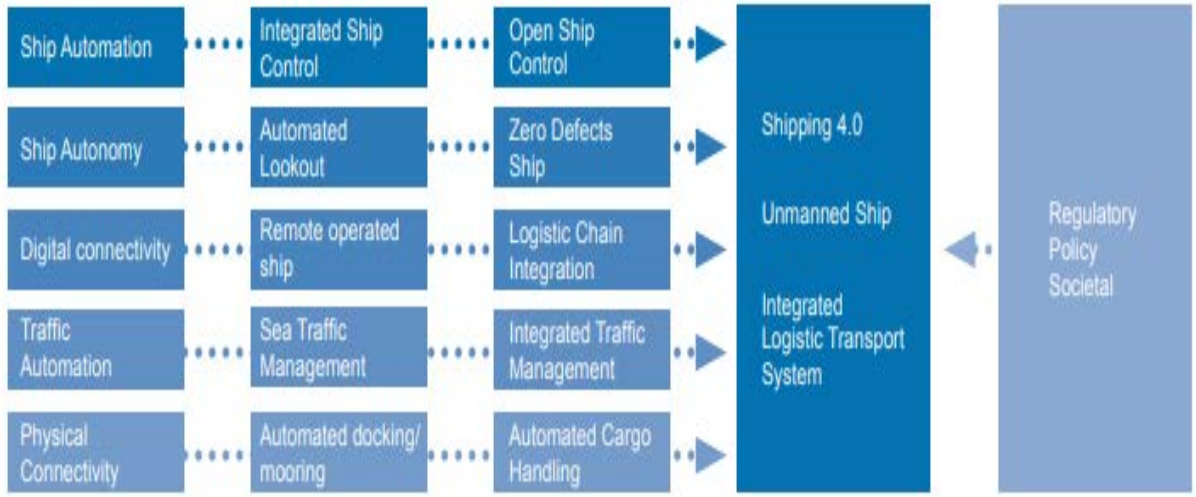


Figure 2. Integrated operational/information exchange Platform/Portal

Integrated operational/information exchange Platform/Portal/Marketplace intends to improve the overall performance of multimodal transport to create a seamless and secure information system. This will be done by interconnecting mobile and wireless communications developments, tracking and tracing, fleet and freight management, and Internet-based technologies. Integrated platforms aim to link all actors together to allow cooperation, collaboration, and information sharing from the point of dispatch to the point of arrival.

Another application towards IoT framework enhancement, the **Fleet Data IoT platform,** draws on different sources for data, including onboard sensors, the ship's voyage data recorder (VDR), or the Integrated Automation System. Data is pre-processed and transferred ashore by satellite connection. Users access a secure online dashboard that is virtually connected via Application Program Interfaces (APIs) to the analytic, monitoring, and management tools available through the IoT.

The platform will also look at integrating edge computing, real-time analytics, artificial intelligence, hyper-precise data, and blockchain. An example application is using sensors incorporated on and in quay walls, dolphins, waterways roads, and traffic signs to generate continuous measurement data and communicate with autonomous systems, laying a path for facilitating autonomous shipping, a target the port aims to meet by 2025. The Port of Rotterdam integrates IoT using sensors (e.g., environmental data) to reduce berthing times by one hour (saving of ~USD $80,000)[4]. **Another area of IoT-enabled technology, Rotterdam's port** invests in is a barge-tracking system known as Port Insight.

[4] https://maritimefairtrade.org/rotterdam-uses-iot-to-save-1-hour-of-berthing-time/

4.6 Conclusion- Big data analytics in Maritime Sector

Big data is considered one of the top initiatives to transform the shipping industry. According to the Global Marine Technology Trends 2030[5] report published in November 2015, big data analytics will be one of the top 18 transformational technologies being used by the sub-sectors (commercial shipping, naval, and ocean) in the marine industry.

The maritime industry is a complex system that requires a rapid adaptation to changing conditions and in which the decision-making process needs to consider many factors.

Big Data analytics tools make it possible to analyze a large quantity of data to gain insight that supports decision-making. However, there are many challenges that this industry must examine, such as cyber security threats, misreporting of data, and lack of cross-enterprise technology implementation.

5. Design and Development of Advanced Data analytics module

5.1 Introduction

Within the SmartShip project, we develop an IoT-based data analytics module. This will be the core of the multi-level optimization of the vessels' operation and management in terms of fuel consumption, energy efficiency, emissions, and circular economy principles.

SmartShip ecosystem comprises three components. From a bottom-up perspective, these components are:

1. Data Sourcing (IoT)
2. SmartShip Core system
3. Users Applications

Data Sourcing (IoT) definition

This component of the SmartShip system is considering tools, communication protocols and network topology for data retrieving, data pre-processing at the edge and finally transferring of information to SmartShip core for further processing and analysis.

In addition, this section refers to the Task 4.3 of SmartShip, which will be responsible for the actual design and development of the IoT-based data analytics module of SmartShip. This module will be the core of the multi-level optimization of the vessels' operation and management in terms of fuel consumption, energy efficiency, emissions and circular economy principles.

SmartShip Core system

The SmartShip core is the heart of the whole ecosystem. Data is processed, analyzed, and visualized to support decision-making for critical maritime operational procedures defined in the project's Use cases.

User Applications

[5] https://www.researchgate.net/publication/297195898_Global_marine_technology_trends_2030

This layer identifies how the meaningful information as product of data processing and analysis is consumed from users either ashore or on board. In this component of the system the user consumes information as represented in SmartShip core, reach decisions for critical tasks (refer to use cases) and configure management of the fleet in an optimized manner in terms of energy efficiency and emission control. This layer also facilitates active interaction between shore and vessel where information is returned to the source (vessel) as valuable feedback for a sustainable and green operation. In this layer is where SmartShip architecture realizes circular economy principals.

5.2 SmartShip Core system Description

SmartShip core system comprises four tiers. Data processed from the source and analyzed to be transformed into meaningful information for supporting users' decision making.

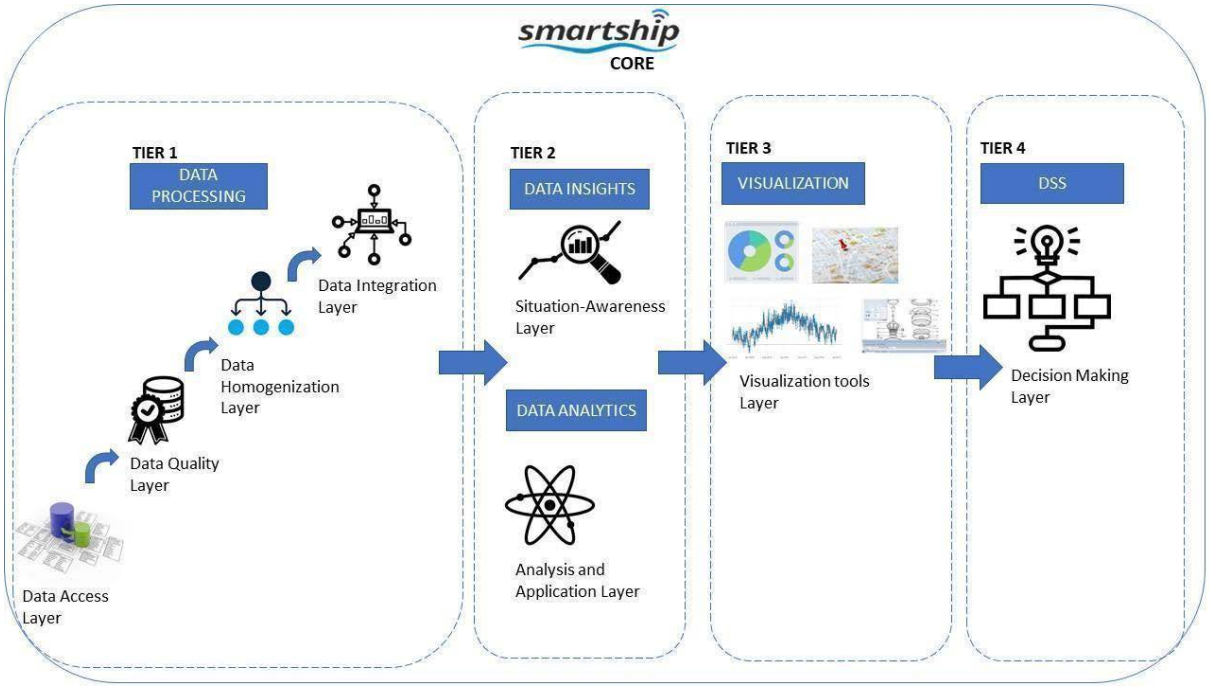


Figure 3. SmartShip Core Tiers

5.2.1 Tier 1. Data processing

Data processing consists of four layers interlinked sequentially. It begins with the data access layer with build-in database views and connectors querying data from a distributed network of data sources. The data quality layer is followed and triggered in two instances (vessel and office sides). The vessel side includes data pre-processing and compression before delivery to shore. The office side is a second step validity, accuracy, and consistency checking once data batches are transferred to shore for further processing. The third layer is related to the data homogenization or conceptual representation layer. This layer deals with the transparent manipulation of data provided by heterogeneous sources and consists of two sub-layers naming the data heterogeneity manipulation and the data uniformity. The homogeneity of multi-source data using ontologies will assist in achieving the correlation of different measurements for the same value from other devices or sources across the fleet. For example, wind speed is streamed from a different acquisition system in vessel A (i.e., Laros) and another third-party provider in Vessel B (ENIRAM). SmartShip is homogenizing these data streams of the same domain acquired from independent parties hiding semantic heterogeneity and enabling common interpretation of data values. Data uniformity sub-layer triggers an auto unit conversion of data pooled from different locales to single measures (e.g., Vessel A gives M/E power in Kw and Vessel B in BHP). The final layer deals with data

integration. This layer associates and integrates the same information from different sources to escalate machine-accessible low-level data to higher-level abstractions suitable for decision making. For example, Vessel A streams fuel consumption for the main engine from a high-frequency flow-meter, telegraphs in daily intervals, and lab analysis after bunkering. SmartShip, through operator-defined rules, combines and synchronizes values to translate data into meaningful information.

5.2.2 Tier 2. Data Insights and Data analytics

This tier of the SmartShip core system deals with the integrated SmartShip advanced data analytics module as a build-in system function. Tier 1 feeds the SmartShip analytics module, where data is presented in two manners:

1. Data insights are the situation awareness data representation field. The user quickly grasps the big picture over large data volumes, observes in real-time, uncovers hidden patterns in the underlying data, and gains knowledge. Data insights could be consumed through the user application interface in nearly real-time close to the edge of the source (fog layer); thus, user onboard will instantly reflect insights to a fast response and decision. Data insights are mostly consumed following the data integration layer (refer to tier 1).

2. Data analytics is the operator-defined algorithmic analysis of fused data. Data analytics are performed ashore (office environment). Analytics allow “hindsight” to reflect and learn from past data by statistical processing past observations (trend analysis, etc.) and detecting hidden correlations among seemingly unrelated data. This is where deep knowledge of various aspects of vessels’ lifecycle is achieved (LCA knowledge base). Analytics gives “insight” interpreting data and responding efficiently to the present by providing KPI's real-time monitoring (operational efficiency, safety performance….), enabling vessel’s benchmarking against theoretical curves. Specifications, tests, sea trials, and competitors or sisters’ vessels trigger timely anomaly detection / alerting for abnormal behaviour and deviation from predefined thresholds. Finally, Analytics offers “foresight” predicting and getting ready for future events by activating what-if scenarios (forecasting based on current observations) and performing risk assessment (multi-factor). Apart from user-defined (based on subject expert judgement) algorithms data analytics module enables machine learning AI models for forecasting.

5.3 **Smartship Data sourcing (IoT) and Advanced Data Analytics modules modelling, configuration and deployment.**

In this section, a demonstration of the configuration of SmartShip Advanced Data analysis module will be presented based on tier 1 and tier 2 of SMARSHIP architecture.

5.3.1 Release Plan

The development of the SMARTSHIP advanced data analysis module followed agile/adaptive methodology with an incremental release of two versions. The first iteration was completed in M36 with a first design/version of the module and the second iteration was concluded in M48. The final release is capitalising on the existing DANAOS infrastructure for data sourcing and processing (DANAOS fleet performance system and DANAOS IoT network on-board) constituting a valuable increment to the system by advancing the functionalities and thus brining benefit to the existing design of vessel performance monitoring.

The implementation team consisting of project researchers (seconded staff) was flexible, self-managed, multi-disciplinary in skills and background and dedicated to the project. The team included a Product

Owner who played the role of the central interface between the users and their requirements as depicted and elicited in D2.1 (WP2) while set priorities in the release backlog prioritising user stories in order of importance. The team supported by "servant leaders/team managers" (Servant Leaders), are consulted the implementation team, coaching and managing the team whilst removed obstacles where raised aiming at the enhancement of team productivity.

5.3.2 Data Sources

On-board IoT

The data sourcing is mainly referring to datasets streamed from vessel data points bundled together in an IoT network on-board. The IoT network along with the main datasets collected from the vessel is thoroughly described in D3.1, section 2.2. In the diagram below is displayed a high-level depiction of the IoT network configured on-board the reference DANAOS vessel that was utilised for the training and the deployment of the SMARTSHIP advanced data analytics module.

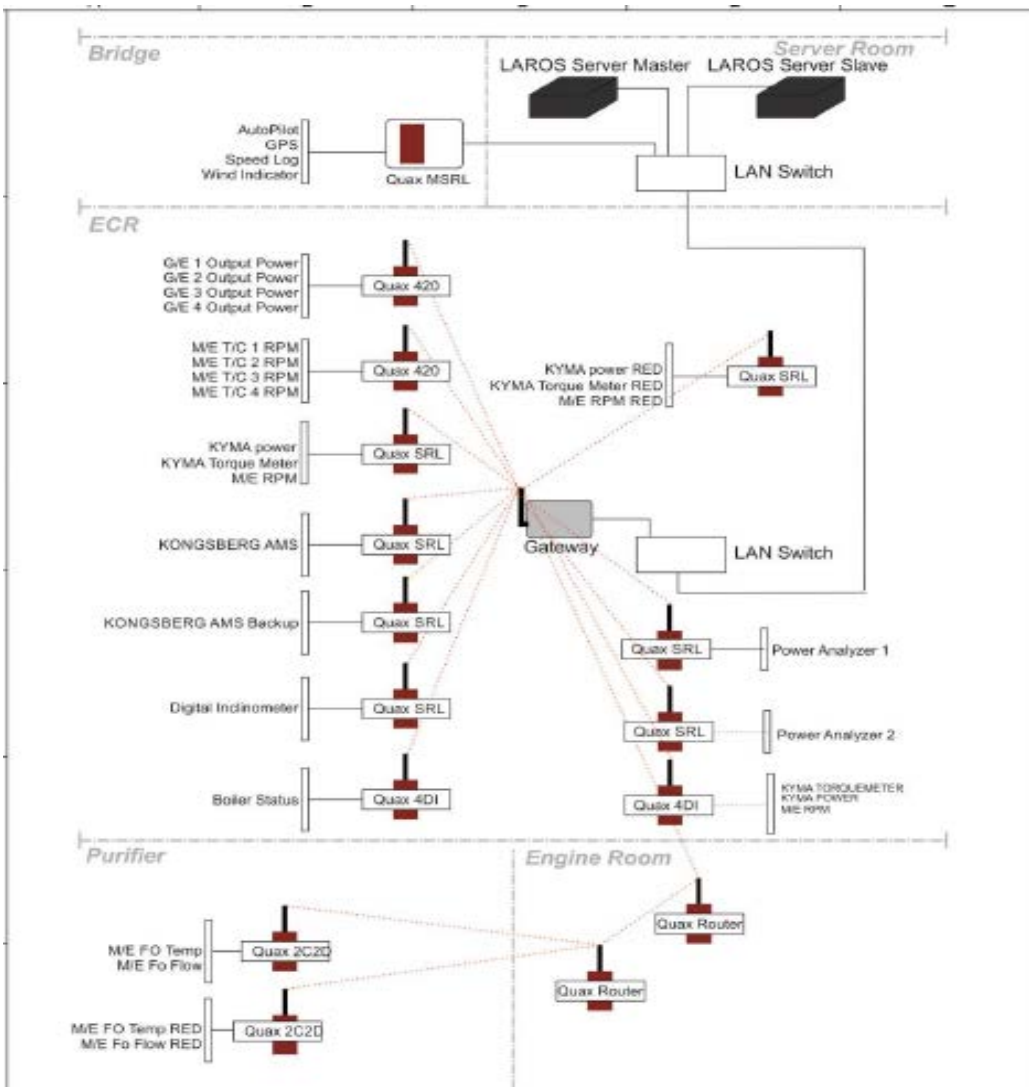


Figure 4. On board sensor suite internal topology & architecture

Data acquisition was also supported and enriched by external databases and services.

Weather API

The weather service API constitutes an endpoint to retrieve the weather state and more specifically the sea state for a specific time and location. The weather is initially acquired from **National Oceanic and Atmospheric Administration** (NOOA), and after appropriately processing is stored in an SQL server in 3-hour intervals. The weather grid has a granularity of 0.5 degrees.

Hindcasts are available up to **2019-09–09**.

- The URI can be found at: https://195.97.37.253:5007/weatherService/latest/lat/lon
- The latest parameter can be replaced with a specific timestamp (e.g: *2022-11-16 09:00:00*).
- Lat, Lon parameters are referring to corresponding latitude and longitude values in decimal degrees.

Snapshot of the stored data as well as the acquired weather features, via the API, in json format is depicted below:

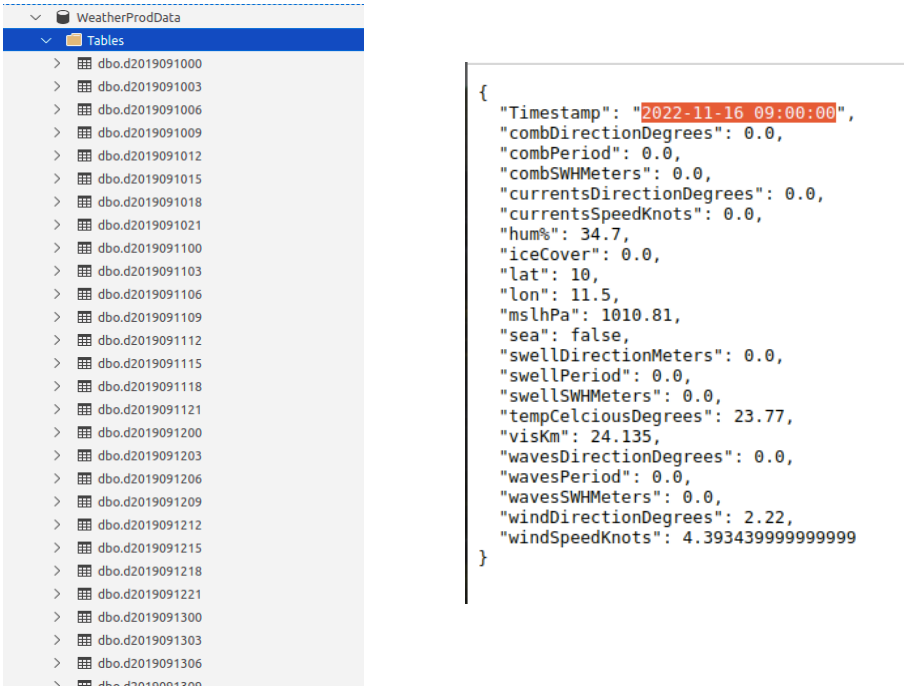


Figure 6: Database snapshot & detailed

The forecasting values for a particular time and location correspond to the closest point of the weather grid, as depicted in the picture below, as well as the closest time in the three-hour interval. We also attaching in Appendix I a code block that represents the implementation of the API interface that is used to acquire the aforementioned weather features.

Figure 6: Requested & Forecasting coordinates from NOOA

Offline Model Training data storage

To store data for the training of the developed Artificial Intelligence and Deep Learning models, the PostGIS[6] database is used. PostGIS is a POSTGRESQL database extension that provides support for geographic objects such as linestrings and points and allows faster query execution and indexing for spatial operations. Therefore, trajectory data from the AIS are directly stored to a PostGIS database and queried from the trajectory classification module either on run-time for real-time monitoring or offline for the training of the developed methodologies.

5.3.3 Condition-based Maintenance Use Case: Application of Voyage fuel consumption and emission monitoring

In the context of SMARTSHIP project we employed a Big Data Analytics - multipurpose - toolkit adapted to the needs of the maritime sector. The proposed framework incorporates a variety of state-of-the-art streaming tools for real-time analysis of vessel data as well as tools for continuous integration/deployment (CI/CD) of ML/DL models regarding operational optimization, causal analysis and event recognition. By utilising DANAOS existing in-house infrastructure concerning Edge-Headquarter (EDGE-HQ) communication between the vessel and the office, we can incorporate the aforementioned pipeline in a broader data acquisition network in order to aggregate, synchronise and process data coming from the vessel in real-time. The resulting platform constitutes a prototype version of a simulation framework, enabling sensing and control actuation on the vessel that aims to assist shipowners to achieve efficiency in fleet management with tangible benefits in terms of emission reduction, environmental compliance and protection of crew safety onboard.

The available resources and necessary requirements in terms of data storing, provisioning and processing were defined in detail by operating an existing Living Lab (PANAMAX Container vessel) as a testbed to realise and validate one of the most prominent use cases for operational optimization and

[6] https://postgis.net/

emission control, namely, ***Fuel Oil Consumption approximation***.

The main building blocks of the proposed BDA framework are described in the following subsections.

Data Analytics Suite: Data Processing – Causal Analysis – Pattern Recognition

Monitoring of the vessel will be performed via the Fleet Performance Monitoring System developed by Danaos acquiring real time measurements concerning the operational state of the vessel. These measurements will be utilised accordingly in order to assess the environmental footprint of the vessel via SOTA data driven predictive schemes for FOC approximation. User defined dashboards and GUIs will offer a holistic insightful representation though tailor made statistical analysis algorithms of the broader range of operational features (e.g. Speed (kn), Power absorbed by the ship propulsion system (kW), Rounds Per Minute of the main Engine (RPM), etc).

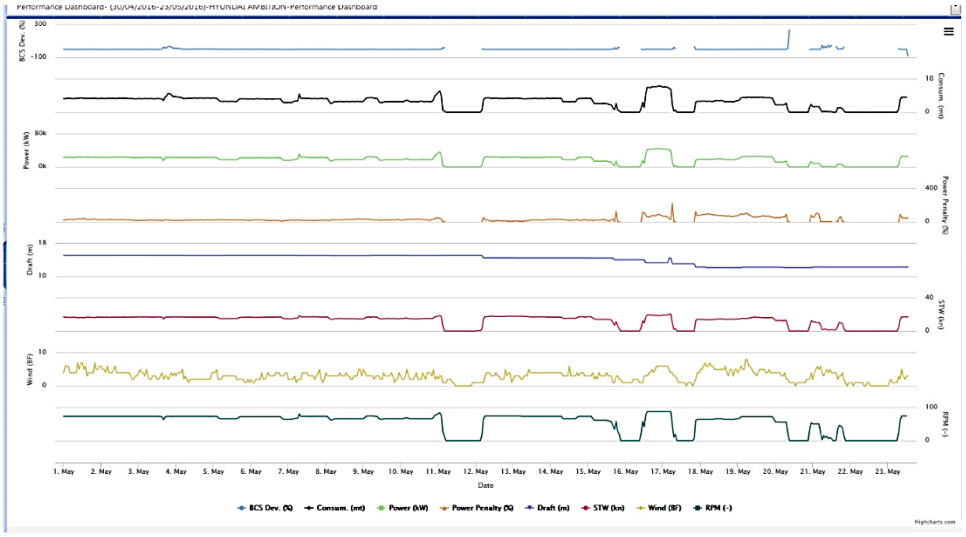


Figure 7. Time Series analysis snapshot

Data coming from on-board sensor instalments concern different compartments of the vessel (Bridge, Engine rooms, Deck) and consists of approximately $\simeq 500$ features. However,
FOC is highly impacted by the total resistance of the vessel as it moves forward, so this hydrodynamic force opposing the movement of the ship along its longitudinal axis, which is known as total resistance, is the most critical feature to estimate in order to correctly predict FOC.

Based on the above, through a streamlined Feature Selection we create a subset of the original set of ~500 features that consists of features which heavily affect total resistance, such as:

- Features that correspond to the frictional resistance and can be utilised in the context of a FOC estimation scheme like Speed Through Water (STW), Draft and Displacement.
- Features that describe the wave resistance components (Wave height/Direction, Wave Period, Swell Wave Height/Direction, Swell Period).
- Features that model the air resistance component (Wind Speed/Direction, Combined Wind Wave Height/Direction, Current Speed/Direction).

In order to further reduce the dimensionality of our dataset, we conduct PCA (Principal Component Analysis) to determine the principal components (expressed as appropriate linear combinations of the initial set of features), that can model the dataset in its entirety, in terms of explainable variance, in the best way possible. As depicted in the following graph (Fig. 3), the variance entailed in the dataset can be attributed, as a whole, to the first 10 components extracted from PCA.

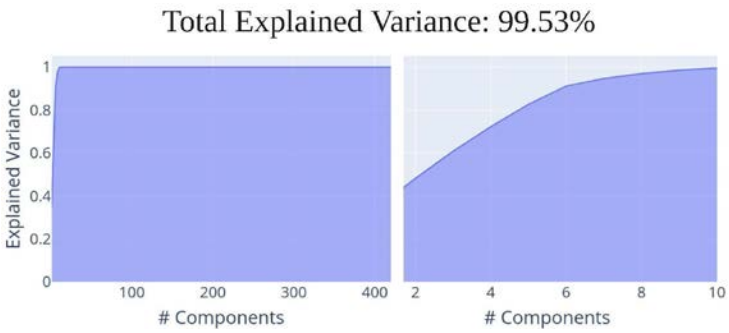


Figure 8. Dimensional reduction via PCA

In the following we demonstrate the applicability of a consolidated approach that combines *Random Forest* and *Spearman Correlation* to extract the most important and independent features to estimate FOC. By applying this algorithm and taking into account the PCA presented above, we conclude with the most important independent features to be exploited in the next sections in order to approximate FOC via data driven methods.

In Table 1 we depict the experimental results from conducting multiple regression analysis utilising the aforementioned algorithm. Detailed description of the features and their abbreviation as well as their measurement unit is also presented.

Table 4. Feature ranking utilising RF algorithm.

	Algorithm		
	RF		
Ranking	**Feature**	Abbrv.	Meas. Unit
1°	Power	P	kW
2°	Speed Through Water	STW	kn
3°	Displacement	Δ	tn
4°	Draft	Dr	m
5°	Combined Wave Height	CWH	m
6°	Swell Wave Height	SWH	m
7°	Current Speed	CS	kn
8°	Current Period	CP	sec
9°	Swell Wave Period	SWP	sec
10°	Current Direction	CD	°
11°	Swell Wave Direction	SWD	°

Data Cleaning

Raw data, collected from the sensors of the vessel, are in time-series (minutely) form and tend to be "noisy" (high variance, high standard deviation from the mean) and in some cases even erroneous. In order to remove noise, Kaklis et al. (2022b) employ a fit & filter technique that effectively "cleans" the data, but at the same time keeps the bulk of information needed for training robust predictive models. The raw vessel's speed and corresponding FOC collected from the sensors versus quasi steady filtered data utilising the algorithm from (Kaklis et al. 2022b) is depicted in **Figure 9.**

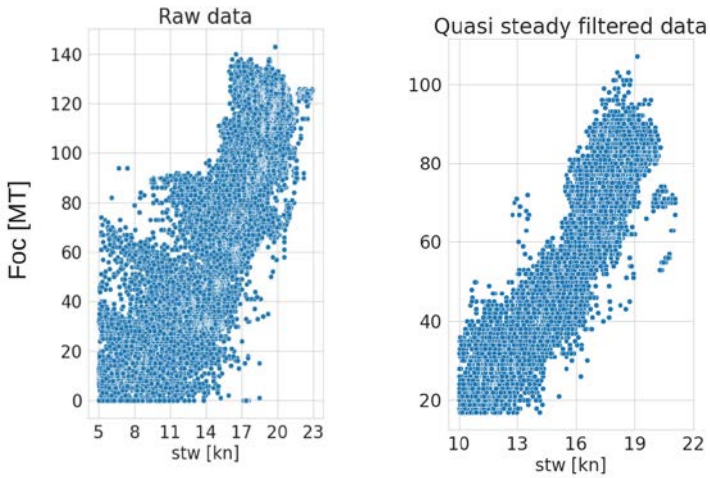


Figure 9. Measured STW vs FOC for a timespan of one year.

An alternative method for removing noise from the dataset is to apply the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) clustering algorithm. DBSCAN is an unsupervised machine learning technique used to identify clusters of varying shape in a data set (Ester et al. 1996). It can identify clusters in large spatial datasets by looking at the local density of the data points. Its main advantage is its robustness against outliers and noise, which are removed from the clustering scheme.

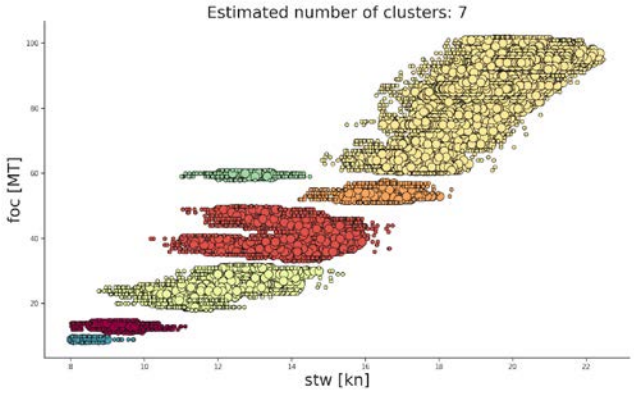


Figure 10. "Cleansed" version of STW vs FOC with DBSCAN

DBSCAN can work without an expected number of clusters (such is the case with the popular K-Means clustering algorithm), and requires two parameters: *epsilon* and *minPoints* to define dense clusters of arbitrary shape. *Epsilon* is the radius of the circle to be created around each data point to check the density and *minPoints* is the minimum number of data points required inside that circle for that data point to be classified as a cluster.

Figure 6 depicts the resulting clusters and corresponds to the points that remain after removing noise. What is even more important, when comparing the plots in Figures 4 (right) and 5 is that the DBSCAN based noise removal method is in agreement with the denoising procedure demonstrated in (Kaklis et al. 2022b). More specifically, if we apply the Kolmogorov-Smirnov (KS) test to the distributions of the two "cleaned" versions of the data, we validate that they are following the same pattern. KS is a non-parametric test (normality is not a prerequisite) that evaluates the maximum absolute difference between the cumulative distributions of the two groups as follows:

$$stat = sup_x|F_1(x) - F_2(x)| \quad (2)$$

where F_1, F_2 are the two cumulative distribution functions and x are the values of the underlying variable (here FOC).
We can visualise the value of the test statistic, by plotting the two cumulative distribution functions and the value of the test statistic as well as their histogram (Figure 11).

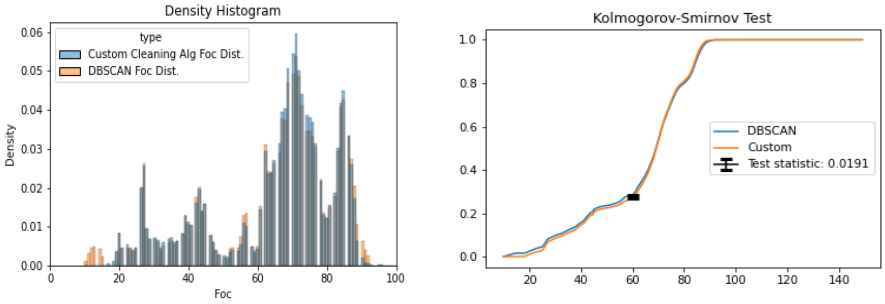


Figure 11. Histogram and KS-Test of the FOC values extracted after applying the Custom de-noising algorithm and DBSCAN

Model Integration/Deployment

The immediate results from the thorough analysis conducted (time series analysis, data cleansing, event recognition, feature selection) in the initial processing layer, are consumed by another building block of the envisaged framework, the Knowledge Hub. The Knowledge Hub (KH) incorporates a variety of multi-disciplinary approaches regarding data provision, re-usability and curation as well as state of the art frameworks for model versioning and deployment. It constitutes a holistic approach that aims to create an adaptive versatile observatory for the shipping industry that comprises structured methodologies for inter-connecting each use case with the appropriate data, processing algorithms and simulation models. All these are joined together adequately, facilitating towards the decarbonization of the maritime sector. Figure 2 illustrates a multi-modal streamlined procedure, stored in KH, adapted to the task of FOC estimation.

The Knowledge Hub aims to largely simplify and standardise the way the various tools and services provided by the DT's ecosystem are operating and communicating with each other, following the standards of an ICT (Information Communication Technology) framework. The general streamlined procedure is based on some functional and non-functional characteristics:

- Data Processing
- Model Configuration
- Model Employment/Deployment
- Decision Support System

Data processing focuses on determining the most important features, spending the use case (here FOC estimation), and data curation accounts for removing the bias (outliers, faulty measurements) from the bulk of data collected in real time from IoT instalments, as already discussed in 5.3.1. As a post-processing step the calculation of correlation between the most important features results in the selection of an ideal feature set that combines importance and independence. The resulting feature set is utilised accordingly in the training process of a data driven FOC estimation model.

In **the figure below,** we depict the general standardised, streamlined procedure adapted to the needs of a specific use case by selecting the appropriate algorithms and models (shown on the side of each customised flow) and applying them into practice.

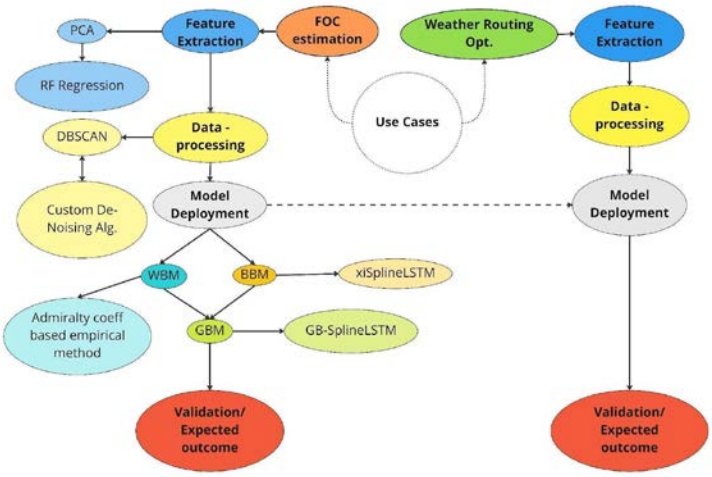


Figure 12. The streamline procedure adapted for the FOC estimation use case.

Furthermore, through a versatile Models Library and Execution Engine incorporating latest technological advancements regarding CI/CD of models (Apache MlFlow, Apache Airflow), Knowledge Hub will be responsible for the appropriate versioning, configuration (framework selection, CPU/GPU optimization, scalability, etc) and deployment (restAPI) of tailor-made mitigation strategies addressing the ultimate goal.

Conclusively, the proposed framework consists of the following components:

- **IoT backbone suite -** *Data acquisition layer*
- **Knowledge Hub -** *Processing - Orchestration - Computing - Deployment layer*
- **Main GUI -** *Visualization layer*
- **Edge Computing -** *Sensing & Control actuation layer - Requirements & Refinements elicitation*

A high-level visualisation of the streamlined procedure demonstrated in previous paragraphs is depicted in Figure 13, and comprises several steps from data collection and filtering to model building, evaluation, selection and deployment as well as the inter-linkage with a DSS described thoroughly in WP5

Data is continuously harvested from different sources (AIS, Sensors, Noon Reports, Weather Service API's) via a state-of-the-art scheduling framework *Apache Airflow*. The pipeline harvests more than 100gb of data on a monthly basis, corresponding to routes of different vessels, which are described by the aforementioned variables. This framework is utilised with the aim to build a fault-tolerant, modular, and multi-purpose big data tool for the maritime industry that is able to harvest data from different sources and perform tasks such as Event Recognition, Causal Analysis, Forecasting, and Incremental Training. In the first steps, the framework integrates streaming algorithms, in *Apache Kafka* and *Spark*, that optimise data collection, processing, and storing. More specifically, the batch streaming process is handled by Kafka Cluster through a centralised distributed messaging system, which allows to balance the load of harvesting data streams in real-time from AIS and on-board monitoring systems. In continuance, the data are processed by exploiting the parallelization capabilities of Apache Spark and are eventually stored in a centralised cloud-based platform. The cleansed version of the data is consumed by a variety of data-driven models that are trained on an ideal feature set for the specified task (e.g FOC

estimation), which has been extracted in the previous step. After training is complete, each model's artefacts (hyper-parameters, training error, evaluation error, convergence plots, size of dataset) are automatically logged in a web-based micro-service (*MLFlow*) to be easily accessible and comparable in order to query the most accurate model in terms of validation error. The selected model is wrapped as a web API service and is queried for inference in real-time from external applications, vendors and/or stakeholders. After selecting the appropriate FOC prediction model, new data streams (i.e., from sensors, AIS) that are pushed to a Kafka topic on a weekly basis, are fetched once a week from the topic and used to update the model. The architecture of this pipeline gives us the advantage to leverage the streaming capabilities of Kafka, the task automation power of Airflow, and the logging features of MLFlow — all structured and orchestrated by a set of Docker containers.

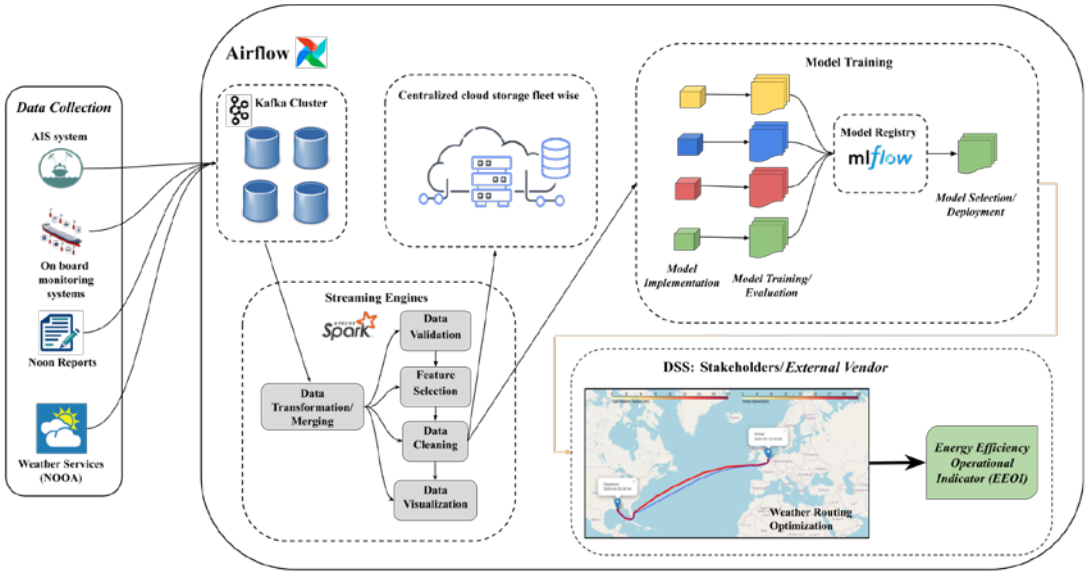


Figure 13. The streamlined procedure from data collection to model deployment.

5.3.4 Application of Smartship Offline training model for trajectories labeling and real-time vessel activity classification based on AIS data.

For the training of the model, representative trajectories of mobility patterns were required to be used as the ground truth. Thus, already labelled trajectories from historical AIS data, which were annotated as “anchored”, “moored”, “underway”, and “fishing”, were used. These trajectories were converted into images, which in turn were used as training instances of the deep learning model. For the implementation, the Keras[7] library with a TensorFlow[8] backend was used, which consisted of not only APIs to create neural networks, but pre-trained CNN models as well. These pre-trained models were employed and fine-tuned to classify images of mobility patterns The Python programming language was used for the training and experimentation.

There are several frameworks for distributed stream processing such as Apache Spark[9], Apache Flink[10], and Kafka streams[11], out of which only Apache Spark has support for the Python programming

[7] https://keras.io/
[8] https://www.tensorflow.org/
[9] https://spark.apache.org/streaming/
[10] https://flink.apache.org/
[11] https://kafka.apache.org/documentation/streams/

language, which was needed for the implementation of the neural networks and the creation of the images. Apache Spark is not preferred since it performs micro-batching over streams of events, and a system is needed that can handle event-processing in real time. Therefore, to balance event-processing with low latency and high throughput, the Apache Kafka[12] framework was used in this phase, a distributed publish-subscribe and message-exchange platform similar to a message queue able to process streams of events as they occur. Three major concepts exist in the Apache Kafka ecosystem, namely topics, producers, and consumers. A Kafka topic is a category/feed name to which messages are stored and published. A producer is an application that continuously publishes or stores messages in a topic. A consumer is an application that is subscribed to a topic and continuously reads or consumes messages. A Kafka topic can be divided into n partitions with each partition storing different messages. Specifically, messages with the same key will be stored in the same partition. n consumers can be subscribed to the partitioned topic with each consumer consuming from a different partition, thus enabling high throughput. A producer can store messages to the partitioned topic, and Apache Kafka will handle the load balancing of the messages among the partitions internally. In our use case, the vessel identifier can be considered as the message key, the AIS receiver as the producer and the trajectory classification modules as the consumers. An even distribution of the load within the nodes of the system reduced the probability that a node would turn into a hotspot, and this property also acted as a safeguard to the system reliability.

The trajectory classification modules were the main components of the methodology. Each trajectory classification module was responsible for consuming AIS messages from a set of vessels and classifying parts of their trajectories based on the deep learning model created in the previous phase. To classify parts of the vessels' trajectories, the module used a temporal sliding window W of user-defined length L and step S. Every S AIS messages, the module took into account all of the AIS messages of the corresponding vessel that belonged to the current window W and converted them to an image. Next, the deep learning model read the image and output for each of the predefined vessel activities a probability. The vessel activity with the highest probability was the final prediction of the module. It is worth noting that in order to leverage execution performance the Scala programming language was used for the real-time trajectory classification modules.

[12] https://kafka.apache.org/

6. Conclusions

This deliverable reports the design and the development of the IoT-based Data Analytics module of SMARTSHIP, which will be the core of the multi-level optimization of the vessels' operation and management in terms of fuel consumption, energy efficiency, emissions, and circular economy principles.

Furthermore, it enables the reader to define the fields of IoT and Advanced Data analytics on how they applied in the maritime sector.

The deliverable is active throughout the Work Package 4 which contains three tasks; Tasks are related with the above aforementioned sections. All the three tasks are ongoing until M48 of the project, in March 2023.

7. References

[1]. V. Kumar, B. Cooper, Z. Cai, G. Eisenhauer and K. Schwan, "Resource-Aware Distributed Stream Management Using Dynamic Overlays," in *25th IEEE International Conference on Distributed Computing Systems (ICDCS'05)*, Columbus, USA, 2005.

[2]. Q. Lin, B. C. Ooi, Z. Wang and C. Yu, "Scalable Distributed Stream Join Processing," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, New York, USA, 2015.

[3]. B. Babcock, M. Datar and R. Motwani, "Load shedding for aggregation queries over data streams," in *20th International Conference on Data Engineering*, Boston, USA, 2004.

[4]. E. Cohen and M. Strauss, "Maintaining Time-Decaying Stream Aggregates," in *Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, New York, USA, 2003.

[5]. B. Butzin, F. Golatowski and D. Timmermann, "Microservices approach for the internet of things," in *2016 IEEE 21st International Conference on Emerging Technologies and Factory Automation (ETFA)*, Berlin, Germany, 2016.

[6]. D. Lu, D. Huang, A. Walenstein and D. Medhi, "A secure microservice framework for iot," in *2017 IEEE Symposium on Service-Oriented System Engineering (SOSE)*, San Francisco, USA, 2017.

[7]. C. L. d. Silva, L. M. Petry and V. Bogorny, "A Survey and Comparison of Trajectory Classification Methods," in *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, Salvador, Brazil, 2019.

[8]. J.-G. Lee, J. Han, X. Li and H. Gonzalez, "Traclass: Trajectory Classification Using Hierarchical Region-Based and Trajectory-Based Clustering," in *The Vldb Endowment - PVLDB*, 2008.

[9]. A. Soleymani, J. Cachat, K. Robinson, S. Dodge, A. Kalueff and R. Weibel, "Integrating cross-scale analysis in the spatial and temporal domains for classification of behavioral movement," *Journal of Spatial Information Science*, vol. 8, 2014.

[10]. L. Sharma, O. Vyas, S. Simon and A. Akasapu, "Nearest Neighbour Classification for Trajectory Data," in *Communications in Computer and Information Science*, 2010.

[11]. S. Dodge, R. Weibel and E. Forootan, "Revealing the physics of movement: Comparing the similarity of movement characteristics of different types of moving objects," *Computers, Environment and Urban Systems*, vol. 33, no. 6, pp. 419-434, 2009.

[12]. C. Ferrero, L. Alvares, W. Zalewski and V. Bogorny, "MOVELETS: Exploring Relevant Subtrajectories for Robust Trajectory Classification," in *33rd ACM/SIGAPP Symposium On Applied Computing*, Pau, France, 2018.

[13]. A. S. J'unior, C. Renso and S. Matwin, "Analytic: An active learning system for trajectory classification," *IEEE computer graphics and applications*, vol. 37, no. 5, pp. 28-39, 2017.

[14]. A. Bolbol, T. Cheng, I. Tsapakis and J. Haworth, "Inferring hybrid transportation modes from sparse GPS data using a moving window SVM classification," *Computers, Environment and Urban Systems*, vol. 36, no. 6, pp. 526-537, 2012.

[15].	S. Dabiri and K. Heaslip, "Inferring transportation modes from GPS trajectories using a convolutional neural network," *Transportation research part C: emerging technologies*, vol. 86, pp. 360-371, 2018.
[16].	Y. Zheng, Q. Li, Y. Chen, X. Xie and W.-Y. Ma, "Understanding mobility based on GPS data," in *Proceedings of the 10th international conference on Ubiquitous computing*, 2008.
[17].	R. Saini, P. Roy and D. Dogra, "A Segmental HMM based Trajectory Classification using Genetic Algorithm," *Expert Systems with Applications*, vol. 93, 2017.
[18].	E. Grimson, X. Wang, G.-W. Ng and K. Ma, "Trajectory Analysis and Semantic Region Modeling Using A Nonparametric Bayesian Model," in *Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2008.
[19].	W. Hu, X. Li, G. Tian, S. J. Maybank and Z. Zhang, "An Incremental DPMM-Based Method for Trajectory Clustering, Modeling, and Retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, pp. 1051-1065, 2013.
[20].	E. Walker and N. Bez, "A pioneer validation of a state-space model of vessel trajectories (VMS) with observers' data," *Ecological Modelling*, vol. 221, pp. 2008-2017, 2010.
[21].	Y. Vermard, E. Rivot, S. Mahévas, P. Marchal and D. Gascuel, "Identifying fishing trip behaviour and estimating fishing effort from VMS data using Bayesian Hidden Markov Models," *Ecological Modelling*, vol. 221, 2010.
[22].	N. Bez, E. Walker, D. Gaertner, J. Rivoirard and P. Gaspar, "Fishing activity of tuna purse seiners estimated from vessel monitoring system (VMS) data," *Canadian Journal of Fisheries and Aquatic Sciences*, vol. 68, pp. 1998-2010, 2011.
[23].	H. Huang, F. Hong, J. Liu, C. Liu, Y. Feng and Z. Guo, "FVID: Fishing Vessel Type Identification Based on VMS Trajectories," *Journal of Ocean University of China*, vol. 18, pp. 403-412, 2019.
[24].	F. Mazzarella, M. Vespe, D. Damalas and G. Osio, "Discovering vessel activities at sea using AIS data: Mapping of fishing footprints," in *FUSION - 17th International Conference on Information Fusion*, 2014.
[25].	A. T. Palma, V. Bogorny, B. Kuijpers and L. O. Alvares, "A Clustering-Based Approach for Discovering Interesting Places in Trajectories," in *Proceedings of the 2008 ACM Symposium on Applied Computing*, Fortaleza, Ceara, Brazil, 2008.
[26].	J. A. M. R. Rocha, V. C. Times, G. Oliveira, L. O. Alvares and V. Bogorny, "DB-SMoT: A direction-based spatio-temporal clustering method," in *2010 5th IEEE International Conference Intelligent Systems*, London, United Kingdom, 2010.
[27].	B. Chuaysi and S. Kiattisin, "Fishing Vessels Behavior Identification for Combating IUU Fishing: Enable Traceability at Sea," in *Wireless Personal Communications*, 2020.
[28].	E. Souza, K. Boerder and B. Worm, "Improving Fishing Pattern Detection from Satellite AIS Using Data Mining and Machine Learning," *PLOS ONE*, vol. 11, 2016.
[29].	X. Jiang, D. Silver, B. Hu and E. Souza, "Fishing Activity Detection from AIS Data Using Autoencoders," in *Proceedings of the 29th Canadian Conference on Artificial Intelligence on Advances in Artificial Intelligence*, 2016.

8. Appendix I

A. Weather service code block snapshot:

```
@app.route("/weatherService/<dt>/<lat>/<lon>")
#@limiter.limit("200/day")
def get(dt, lat, lon):

    try:

        lat = float(lat)
        lon = float(lon)
        ##lon AND lat SHOULD BE PROVIDED WITHOUT ROUNDING
        ##FORM WEATHER HISTORY DB NAME TO SEARCH
        dt = dt if dt != 'latest' else datetime.now().strftime("%Y-%m-%d %H:%M:%S")
        date = str(dt).split(" ")[0]
        print(date)
        day = int(date.split("-")[2])
        month = date.split("-")[1]
        year = date.split("-")[0]
        hhMM = str(dt).split(" ")[1]
        h = int(hhMM.split(":")[0])
        M = int(hhMM.split(":")[1])
        if int(M) > 30:
            h += 1
        if h % 3 == 2:
            h += 1
        else:
            h = h - h % 3
        if (h == 24):
            if (month == '02'):  # February
                if (int(year) % 4 == 0):
                    if (day == 29):
                        month = int(month) + 1
                        day = 1
                    else:
                        day = day + 1
                else:
                    if (day == 28):
                        month = int(month) + 1
```