A Big Data Architecture for Managing Oceans of Data and Maritime Applications

Ioanna Lytra*, Maria-Esther Vidal†, Fabrizio Orlandi*, Judie Attard*

*Enterprise Information Systems
University of Bonn & Fraunhofer IAIS
Germany
† Universidad Simón Bolívar
Venezuela
{lytra,vidal,orlandi,attard} @cs.uni-bonn.de

Abstract—Data in the maritime domain is growing at an unprecedented rate, e.g., terabytes of oceanographic data are collected every month, and petabytes of data are already publicly available. Big data from heterogeneous sources such as sensors, buoys, vessels, and satellites could potentially fuel a large number of interesting applications for environmental protection, security, fault prediction, shipping routes optimization, and energy production. However, because of several challenges related to big data and the high heterogeneity of the data sources, such applications are still underdeveloped and fragmented. In this paper, we analyze challenges and requirements related to big maritime data applications and propose a scalable data management solution. A big data architecture meeting these requirements is described, and examples of its implementation in concrete scenarios are provided. The related data value chain and use cases in the context of a European project, BigDataOcean, are also described.

Keywords—Big Data; Big Data Applications; Oceanographic Data; Maritime Applications

I. INTRODUCTION

Big data technologies are currently being applied to an increasing number of domains and disciplines, proving to be one of the key technological trends of the decade. Big data has been widely and successfully adopted in several domains such as health care, retail, and manufacturing. However, Big data applications in the maritime domain are still underdeveloped. Especially considering that seas, oceans, and other marine areas cover almost 72% of our planet, and about 95% of this realm is currently largely unexplored. Marine areas are one of the most exploited “economic platforms” of mankind; for example, through the fishing industry, boosting the economic and strategic growth of coastal regions (such as through tourism, transportation, and logistics), and also through being a source of renewable energy. Relevant studies evince that the “blue economy”, although difficult to monetize, is clearly a significant and strategic industry and has the potential to grow even further. Moreover, the maritime domain has recently started to offer a wide variety of large and heterogeneous data sources. Sensor devices on all kinds of ships and boats are becoming increasingly common. Apart from GPS sensors for positioning and navigation, the shipping industry integrates also other kinds of sensors to their products, including equipment that provides flows of ship performance, condition data, temperature and humidity sensors. Additionally, energy wasting meters, as well as devices such as fish finder radars or other sonar-based hardware, is becoming increasingly common on vessels. From an environmental perspective, other types of sensors are also being deployed in the sea, mostly for scientific reasons. Sea sensors mainly come in the form of surface or even sub-merged sea-buoys, measuring temperature, sea streams, wind, luminance, sea and wave levels, salt concentration, water pollution, ice movements, and tsunami indications, just to mention a few.

All aforementioned data sources complement each other and can be currently analyzed “offline” in a fragmented way, depriving stakeholders of the ability to exploit their full potential. From a technological point of view, big data comes with common recognized challenges; in many cases known as the 5 Vs of big data, namely Volume, Variety, Velocity, Value, and Veracity. Carefully designed systems are needed in order to be able to integrate and analyze heterogeneous data coming from different sources. These systems should be cloud-based, in order to ensure seamless and universal data access. Moreover, the paradigm of high-performance computing (HPC) plays a very important role, especially when real-time data processing is needed. Interesting cases of such systems (an example is depicted in Fig. 1) for sensors, buoys, and even satellite data exploitation can be found; however, cross-sectorial and cross-domain applications are still missing.

In this paper, we propose the first big data architecture for multi-segment maritime applications that combine data of different velocity, variety, and volume under an inter-linked, trusted, multilingual engine. Such architecture aims at offering

1http://www.forbes.com/sites/eric.savitz/2012/10/22/gartner-10-critical-tech-trends-for-the-next-five-years/
2https://www.forbes.com/sites/bernardmarr/2016/08/25/the-most-practical-big-data-use-cases-of-2016
3http://ieeexplore.ieee.org/servlet/opac?punumber=6687317
4http://internetofthingsagenda.techtarget.com/feature/Maritime-company cuts-through-the-noise-of-sensor-data-analytics
5http://oceanservice.noaa.gov/facts/exploration.html
a big data repository of value and veracity to its end users, i.e., stakeholders involved in maritime applications. Ship builders, ship owners and managers, freight forwarders and logistics companies, tug operators, charterers and cargo consignees, public authorities (e.g., coast guard, border control, police and maritime security authorities), and NGOs (e.g., search and rescue authorities) are just some examples of potential stakeholders. Following an analysis of the main challenges and requirements for big maritime data applications (Section II), the envisioned big maritime data architecture is presented in Section III. As a concrete application of this architecture, in Section IV, references to the use cases in the BigDataOcean European project are explained. BigDataOcean is supported by four different pilots, led by organizations in the maritime domain; pilots are related to vessel’s fault prediction and maintenance, environmental protection, security and detection of anomalies in ship routes, and production of electrical power through wave energy. The proposed architecture satisfies the requirements of all the aforementioned pilot applications. In Section V, the BigDataOcean data value chain is defined and illustrated with concrete examples of data management and data analytics in real concrete scenarios related to the project. Finally, related work is discussed in Section VI, and Section VII concludes and presents our future work plan.

II. BIG MARITIME DATA CHALLENGES AND REQUIREMENTS

The maritime domain encloses large data sources that include data about ships, oceans, wave stations, observations from environmental conditions, fishing and maritime biodiversity, routes and trajectories, and incidental or voluntary oil spill events. At the level of ships and following regulations established on Chapter V of SOLAS, ships on all voyages should collect and store large volume of in-situ data describing real-time conditions of a ship, e.g., motion, position, speed, order and response of engines and rudders, as well as terrestrial and space radio communications and meteorological conditions. Ocean conditions are also frequently monitored and streamed. Longitudinal data sources describe marine monitoring and forecasting in terms of parameters such as currents, temperature, salinity, sea ice, sea level, wind, and biogeochemistry. Scientific and governmental organizations like NOAA, NASA, and NODC provide integrated Web infrastructures to make available in-situ and remote sensor marine data, metadata, and products. Finally, maritime companies offer Web services to access live maps, and historical data about vessels,
voyages, and ports. The very nature of maritime data imposes challenges that need to be addressed in order to facilitate the development of methods to generate and exploit knowledge to sustainably process and analyze maritime data. In this section, we describe maritime data in terms of the 5Vs of Big Data, as well as challenges and requirements necessary to provide reliable basis for users to transform maritime data into useful operational products and findings.

A. The 5Vs of Big Maritime Data

In a general sense, big data is defined as data whose volume, acquisition speed, data representation, veracity, and potential value overcome the capacity of traditional data management systems [1]. Big data is characterized by a 5Vs model: Volume denotes that generation and collection of data are produced at increasingly big scales. Velocity represents that data is rapidly and timely generated and collected. Variety indicates heterogeneity in data types, formats, structuredness, and data generation scale. Veracity refers to noise and quality issues in the data. Finally, value denotes the value that can be obtained from processing and mining big data. According to this 5Vs model, maritime data is characterized as follows:

- **Volume:** Existing maritime data sources provide enormous quantity of data. Public websites from scientific organizations like NOAA, NASA, and NODC collect terabytes of data per day, and allow for accessing a large number of oceanographic and maritime biodiversity datasets. Furthermore, very large databases describing live and historical ship conditions and trajectories are accessible in different maritime company websites. For example, Fig. 1 presents a Standard map produced by a maritime company named MarineTraffic\(^{13}\), reporting location of vessels, tankers, crafts, and fishing ships on April 26, 2017 at 22:49 CEST. Data used to perform this type of live maps comes from satellites and marine databases with more than 18 million vessel and port related events recorded monthly, and 800 million vessel position recorded monthly.

- **Velocity:** Existing maritime regulations establish that ships should collect and store large volume of data describing their ship live conditions, e.g., location, speed, trajectories, and communication audio and engine, rudder, and thruster status. Data recording should be continuous and retained in the ships devices for decision making and analytics. Similarly, organizations like NOAA offer real-time about meteorological and oceanographic conditions collected (half) hourly via the Geostationary Operational Environmental Satellites (GOES) and commercial satellites. Only the NOAA National Center for Environmental Information (NCEI) houses over 20 petabytes of atmospheric, coastal, oceanic, paleoclimatological, and geophysical data, and archives over 26 terabytes of oceanic, atmospheric, and geophysical data each month\(^{14}\). Moreover, maritime companies make available live maps reporting trajectories and location of more than half million vessels all over the world.

- **Variety:** Maritime data is collected in a wide variety of ways and stored in many different formats and data structures. Ship data is collected by tens of different devices and in different formats, e.g., in-situ automatic identification systems (AISs) receive data about locations from navigational satellite systems, collect information from other ships or ports through radio communication systems, while a ship engine performance is monitored by several sensors. In a similar way, oceanographic data is collected using different devices. For instance, time-series measurements at a fixed location can be collected using instrumentation arrays, while other types of equipment, e.g., surface drifters or gliders can freely move and gather oceanographic data. Moreover, sensors affixed to orbiting satellites can continuously gather data about the ocean.

- **Veracity:** Marine measurement precision has been increased in modern devices. However, because data is collected following a wide variety of methods and using disparate instruments, different degrees of accuracy and uncertainty can be observed in maritime data sources. For example, untrained sailors may utilize poorly-calibrated and low-resolution instruments on board ships, or oceanographic buoy measurements may be impacted by buoy size, shape, ballast, and mooring. Thus, missing or inconsistent observations may be reported, and ambiguities may be present across maritime data sources. In order to achieve veracity of Big maritime data, quality assessment methods like deduplication, disambiguation, and domain specific data cleaning process are required.

- **Value:** Big maritime data can be utilized and exploited to enhance the value of diverse maritime data-driven applications. For example, ship and meteorological data represent the basis for the prediction of technical operation and maintenance of vessels and stations. Furthermore, historical ship data can be used to drive energy efficiency improvements, while routes and trajectories can allow for anomaly detection and safety performance. Finally, management and monitoring of accident and environmental risks can be also achieved by analyzing vessel routes and trajectories, historical and live ship and meteorological data, as well as biodiversity data.

Fig. 2 summarizes the very nature of Big Maritime Data.

B. Requirements for Big Maritime Data Applications

The development of maritime applications for data processing and analytics are challenging because of the very nature of the maritime data, i.e., increasing large data volumes, wide range of data types, and different data quality issues. Fig. 3 illustrates the steps that compose the maritime data value chain to be implemented in Big maritime data applications.

First during data acquisition, diverse methods are required to collect raw maritime data gathered from different devices and instruments, and reported in various formats; thus, variety is addressed. Once maritime data is collected, diverse data

\(^{13}\)https://www.marinetraffic.com/en/ais

\(^{14}\)https://www.ncei.noaa.gov/
5 Vs of Big Maritime Data

Fig. 2: Big Maritime Data. Big Maritime data is characterized by 5 properties: Velocity, Variety, Volume, Value, and Veracity: (i) Datasets collected at different Velocity, e.g., real-time observations of weather or environmental conditions, (ii) There is a wide Variety of data published across disparate datasets, e.g., historical ship maintenance data, streaming weather data, and open satellite data, (iii) Maritime Data includes large Volume datasets, i.e., petabytes of data collected by scientific, research, and maritime companies, (iv) Data processing and mining tasks on top of Big Maritime data allow for adding Value, e.g., predicting ship maintenance, modeling oil spill, detecting anomalous ship trajectories, (v) Datasets may suffer of different data quality issues and quality assessment is required to enhance Veracity, e.g., entity deduplication and disambiguation.

analytic methods are necessary to determine accuracy and data integration problems, thus tackling veracity. During data curation, quality issues are assessed, and data is also integrated. Then, different methods for data compression, indexing, and accessing are required to efficiently store Big maritime data. Finally, data processing and analytics are necessary to identify patterns and predict relations across maritime data, thus allowing for value enhancing. Moreover, during all the steps of the data value chain, privacy issues, data protection, and security risks should be addressed. Additionally, tools for effective utilization, analysis, and visualization of Big maritime data need to be available; these tools need to be able to support both stream and heterogeneous data. Finally, data governance and provenance both have to be taken into account in every step of the data value chain. Both data management and analytics techniques need to scale up to large datasets and be able to manage data produced at different speeds, i.e., tackle volume and velocity characteristics of Big Maritime Data.

III. BIG MARITIME DATA ARCHITECTURE

In this section, we introduce a Big Maritime Data Architecture for addressing the challenges and fulfilling the requirements listed in Section II. The goal of this architecture is to efficiently handle big, constantly growing data coming from heterogeneous sources in real or almost-real time, while enhancing and integrating data using semantic technologies.
The Big Maritime Data architecture, depicted in Fig. 4, relies on a Data Lake to store both streamed data generated by sensors, as well as historic, social, scientific, biological, and open data. In the Data Lake, data is ingested in raw formats, e.g., CSV, JSON, or RDF, and can be accessed using native data management engines, e.g., SQL or NoSQL engines. Controlled vocabularies and ontologies are utilized to represent and express the data meaning and main properties of the underlying data. The bottom layer of the platform is responsible for integrating, curating, and semantically enriching data in the Data Lake on-demand, in an efficient and scalable way. On top of this layer, query processing, analytics, and visualization modules are provided for further processing and analysis of the underlying data. A novel query processing engine offers query execution methods able to manage disparate maritime data. In both layers, privacy, trust, security, and access control policies are enforced. Finally, the Big Maritime Data architecture provides applications and APIs to multiple stakeholders with different characteristics and needs. Examples of possible stakeholders include public authorities, entrepreneurs, maritime companies, NGOs, local communities, scientists, and energy companies. These stakeholders can act as data providers and data consumers, as well as users of value added services, ranging from analytics and data integration, to data visualization and query processing. In the following paragraphs, we get into the details of the Data Lake and the big maritime data modules of the architecture.

a) **Data Lake**: The data lake provides a central repository for raw maritime data – unstructured, semi-structured, or structured – that may be available in different formats. This load-first paradigm addresses in the first place the problem of disconnected information silos, as a result of heterogeneous non-integrated data sources in isolated repositories. The data lake allows for large-scale and data stream management; thus addressing volume, velocity, and variety.

b) **Big Maritime Data Modules**: The Big Maritime Data architecture comprises the following modules. Very characteristics of the Big Maritime Data are addressed by each module.

- **Semantic Enrichment**: The semantic enrichment module aims at describing data meaning while decreasing the storage overhead of the produced data. It targets both static and streaming data and can be applied both before the data ingestion, as well as on demand. In the semantic enrichment process, a number of vocabularies and ontologies from the maritime domain can be reused and extended. Examples of such ontologies are included in the (1) Natural Environment Research Council (NERC) Vocabulary Server\(^\text{15}\), (2) Marine Metadata Interoperability repository\(^\text{16}\), (3) Biological and Chemical Oceanography Data Management Office BCO-DMO vocabularies\(^\text{17}\), and (4) Rolling Deck to Repository (R2R) vocabularies\(^\text{18}\). In addition, other general-purpose ontologies can be used to provide meaning to disparate maritime data. For instance, the semantic sensor network ontology (SSN)\(^\text{19}\) is a general-purpose ontology to describe sensor devices, their capabilities, observations, and other sensor-related concepts, and can be leveraged in order to semantify sensor data coming from vessels or buoys. The main goal addressed by the semantic enrichment module is to provide semantic annotation of Big Maritime Data without negatively impacting data volume. Only data required for query answering will be annotated; thus, no materialization of the full datasets will be performed.

- **Data Curation**: The Data Curation includes assessing data quality as well as data filtering, validation, and transformation. The main goal of this module is to improve the accessibility and quality of data, and to ensure that data is trustworthy and accessible, and eventually, to make this data available for use by the other modules. Data curation can be conducted both off-line or on-demand during query processing.

- **Data Integration**: The Data Integration module provides the tools for linking data from a variety of sources. This module enables linking existing repositories of open scientific relevant data with data coming from sensor streams, and other directly or indirectly related sources of data, in order to increase data interoperability. Data integration and homogenization will be performed on-demand during the process of semantic enrichment.

- **Query Processing**: The query processing module provides a common query interface for querying heterogeneous data and metadata that are stored in raw format in the data lake. Query answers can be consumed by either users or other modules for further analysis, like the Data Analytics and the Data Visualization modules. The main novelty of the query engine is the capability of performing a federated query processing over heterogeneous sources stored in raw format in the data lake; the engine is also able to interoperate among the native query engines of these sources. Query decomposition and optimization techniques are tailored for exploiting metadata about the data in the data lake; thus data heterogeneity is overcome. Furthermore, physical operators allow for the adaptation of query execution schedulers to the heterogeneous con-
Heterogeneous data sources are kept into a data lake in raw format and are accessed using native data management engines. Semantic enrichment, data curation, and data integration are performed on top of the data lake on-demand. A query processing engine executes queries over the data lake. Data analytics and visualization are also performed against the data lake. Access control, security, and privacy policies are enforced whenever data is accessed in the Big Maritime Data Architecture.

- **Data Analytics**: This module provides data analytics and knowledge extraction features, including classification, clustering, and segmentation algorithms. Data analytics are applied both on near-real time, as well as on historic data from the maritime domain. The use of the Data Analytics module enables various types of maritime stakeholders to utilize and share analytic methods on big data for the discovery of new knowledge and patterns.

- **Data Visualization**: The Data Visualization module provides users with information about and insights into the maritime data, and assist them to discover the inherent structure, relations, and patterns of their data. Data visualizations help users comprehend big amounts of data and their interrelations as well as improve decision making.

- **Security**: Connection authentication (e.g., SSL, HTTPS) and security methods (e.g., MD5 hashing) ensure a secure data infrastructure and exchange.

- **Privacy**: The Privacy module ensures that privacy-sensitive data, e.g., data coming from proprietary datasets that may include personal information, is anonymized before it is exposed to the users.

- **Access Control**: The Access Control module allows, for instance, private or business collaborators to have access to provide and consume private datasets.

**IV. APPLICATIONS IN BIGDATAOCEAN PROJECT**

BigDataOcean is a 30-month H2020-RIA project started in January 2017, targeting maritime big data applications for EU-based companies, organizations, and scientists (http://www.bigdataocean.eu). Its main objective is to deliver services for various stakeholders on top of a multi-segment platform that will address the velocity, variety, and volume of maritime-related data and provide an inter-linked, trusted, and multilingual engine. In particular, BigDataOcean aims at leveraging existing modern technological breakthroughs in the areas of Big Data and Linked Data, and rolling out a completely new value chain of interrelated data streams coming from diverse sectors and heterogeneous sources. We present four pilots in...
the context of BigDataOcean. Main challenges and role of the Big Maritime Data architecture are also described.

A. The BigDataOcean Pilots

The BigDataOcean project is supported by four different pilots led by organizations and companies in the maritime domain. The pilots are related to vessel’s fault prediction and maintenance, to environmental protection of seas, to security and detection of anomalies in ship routes, and to production of electrical power through wave energy. As can be viewed in Fig. 5, the four pilots will act both as data providers, and data and service consumers.

1) Fault Prediction and Proactive Maintenance:

**Pilot Owners:** The pilot owners are two Greek shipping companies ANEK and FOINIKAS.

**Challenges:** Unpredicted damages and mechanical failures increase the costs of shipping companies, since they are related with high costs of repairs and spare parts, loss of earnings due to immobilization of vessels, and probable environmental damages. To minimize such incidents, contemporary vessels are equipped with sensors and monitoring utilities for collecting data. Processes and systems in this domain are vessel-centric and the full potential of the sensor data along with related historical and external data has not been exploited yet.

**BigDataOcean Platform:** The pilot owners will provide anonymized vessel data and consume weather data, historical data based on incidents, buoys data, such as water speed and temperature, and feedback received from vessel’s crew and domain experts. Eventually, they will take advantage of predictions and analytics by the BigDataOcean platform in order to perform proactive maintenance, and thus reduce costs and unwanted damages.

2) Mare Protection:

**Pilot Owners:** The pilot owner is the Hellenic Centre of Marine Research (HCMR).

**Challenges:** Oil spill models provide contingency planning and effective response strategies against hazardous oil spills at sea in case of ship accidents. The reliability of such models is of crucial importance since their results are widely used to support operational activities in the marine environment. However, the integration of related systems and data in these simulations with the aim of increasing the forecasting skills on the marine and environmental impact of oil spills is still open.

**BigDataOcean Platform:** Cross-sectorial data stored in the BigDataOcean platform will be ingested into numerical models in order to improve their accuracy. Examples include weather data, data from coastal or offshore stations about atmospheric, hydrodynamic, and sea state conditions, satellite data, pollution reports, and AIS (Automatic Identification System) data. In addition, HCMR can provide results from its hydrodynamical models to the BigDataOcean repository that can be reused by other stakeholders in different use case scenarios.

3) Maritime Security and Anomaly Detection:

**Pilot Owners:** EXMILE, initiator or MarineTraffic, a community-based project that provides real-time information about vessel movements and port traffic, is the main pilot owner.

**Challenges:** The Maritime Domain Awareness includes the understanding of activities, events, and threats in the maritime environment, having direct impact on security, economic activity, or the environment. The understanding of the complex maritime environment requires the ability to identify patterns from big amounts of data from monitoring thousands of vessels. The main challenge is to combine the data fused from various sources and generated by vessel sensors daily in an efficient and effective way, in order to act proactively and, thus, minimize the impact of threats in the sea.

**BigDataOcean Platform:** The integration of cross-sectoral data, such as weather and nautical information and incident reports, in combination with machine learning algorithms, can assist in improving the predictions and classify “anomalies” related to terrorism, illegal trafficking, fishing, etc. EXMILE provides the BigDataOcean repository with vessel tracking data from more than 460,000 vessels that can be exploited in the context of the other pilots as well as other stakeholders.

4) Wave Power:

**Pilot Owners:** The owner of this pilot is NESTER, the responsible company for the energy transport infrastructure in Portugal.

**Challenges:** Wave power can contribute with massive amounts to the overall energy picture. However, the wave energy industry is still in its infancy due to various technological hurdles that need to be overcome. Predicting the best locations, the expected energy production and equipment costs, as well as the environmental impact on the marine life is very challenging and requires accurate data aggregation.

**BigDataOcean Platform:** The BigDataOcean platform will enable the integration of existing data, e.g., environmental and geophysical with other marine data coming from different sectors, e.g., vessel position data and offer improved analytics; therefore, contributing to the wave energy studies. Moreover, wave and tide data generated by the pilot owner, may be useful not only for other energy companies but also for scientists, NGOs, and maritime companies.

V. DATA VALUE CHAIN IN BIGDATAOCEAN

In order to ensure that the BigDataOcean platform is capable of aggregating and synchronizing the requirements from the
Big Maritime Data Architecture

1. Fault Prediction
2. Mare Protection
3. Anomaly Detection
4. Wave Power

Fig. 5: The BigDataOcean Pilots. Four pilots support the BigDataOcean project. (1) Fault Prediction and Proactive Maintenance: predictions and analytics provided by the BigDataOcean platform will allow for conducting ship proactive maintenance. (2) Mare Protection: Mathematical models will be enhanced with data available in the BigDataOcean platform to improve the prediction of the environmental impact of oil spills. (3) Maritime Security and Anomaly Detection: integrated and semantically enriched data will support the identification of illegal activities, e.g., terrorism, illegal trafficking and fishing. (4) Wave Power: methods to identify the best wave station locations will be implemented on top of the BigDataOcean platform.

By leading these five phases on the four pilots defined in Section IV-A, we are therefore able to concretely identify the requirements that the BigDataOcean platform should satisfy. We here use the third pilot, Maritime Security and Anomaly Detection, as an example. We start by leading out the first phase: the Preparation phase. Through the use of a number of forms shared with the various actors within this pilot we identify the different participating stakeholders, the various activities carried out, as well as the used data sources. Moreover, this information enable us to identify the Data Value Chain occurring within this pilot. Therefore, in the case of the third pilot, we identify the main stakeholders to be the following:

- Port Authorities - Ports are public business entities that are in charge of operating ports and other transport infrastructures. Provided services include enforcing existing port policies, coordinating with the various General State Administration bodies, training, and management.
- Ocean Observatories - Such observatories gather oceanographic data and model services and technologies to support oceanography and marine and coastal research.
- Port/Cargo Community Systems - Complex information systems that optimize logistics processes, particularly within international trade hubs such as ports and airports, by enabling data flows between the various stakeholders.
- Transport and Logistics - Software technology providers

Fig. 6: Example of the Data Value Chain. The Pilot 3 of the BigDataOcean project is used to define the data value chain presented in Fig. 3. Vessel tracking data is ingested into the data lake. Vessels in the Aegean Sea are curated and filtered, as well as data related to these vessels. Then, vessels are classified by type and stored in a database in the data lake. A service for vessel monitoring is then deployed.
develop solutions for the Maritime Industry, including software solutions for security, transport logistics, etc.

- Harbour Pilots and Maritime Consultants - These are mostly concerned with managing (on-board) the movement of vessels in a harbour environment.

After identifying the stakeholders within this pilot, we proceed to identify the sources of used data. These include:

- National Port Statistics;
- Proprietary AIS Terrestrial Network;
- ATO⁄N installations;
- Observing facilities (research vessel, HF radar, glider, fixed stations, beach monitoring, satellite);
- Numerical models (e.g. high-resolution ocean circulation predictions; wave forecasts; extreme sea level oscillations predictions, Metocean parameters);
- Vessel Monitoring System;
- Logistics Data;
- Commercial Data;
- Administrative Data;
- Vessel Tracker (Vessel Position and routing, Weather, Port information);
- Shipping Information Pipeling (Vessel Position);
- Vessel information and tracking data (from AIS data providers).

Based on this aggregated knowledge, as well as the identified processes carried out within this pilot, the Data Value Chain, shown in Fig. 6 is defined. This Data Value Chain shows how the stakeholders exploit the data mentioned above to achieve their aim of anomaly detection on vessels.

The next phase is the Elicitation phase. The aim is here to identify the raw requirements that need to be met. Therefore, in this phase workshops are held for each pilot, where the involved stakeholders describe their needs in managing and using maritime big data. For example, Pilot 4 is associated with various requirements, including the availability of a number of datasets, the possibility of correlation and interpolation, and the availability of filtering services. Further, there are more requirements that are pilot-specific, including but not limited to:

- Provide descriptive, predictive, and prescriptive analytics related to the status of vessels calling ports and their movement tracking;
- Visualization of events related to the vessels’ calling ports such as “on schedule”, “delayed”, or “arrived”;
- Provide information exchange and in-time accurate alerts on movement anomalies of vessels through visualization tool;
- Provide data analytics related to vessels’ movement tracking in cases of anomaly detection to the user; and
- Provide notification alerts related to vessel position, weather/event/incidents.

The third stage of the methodology being used is the Analysis of the requirements previously elicited. Typically, these requirements are unorganized and therefore it is quite challenging to forward such requirements to the technical teams, who would need a certain amount of detail. This phase hence takes care of categorizing the requirements by types (e.g., functional or business), data sources, or domain. After being categorized, a refinement activity removes any existing repetition, and clarifies the requirements to make them unambiguous, complete, consistent, verifiable, traceable, relevant, and feasible.

The fourth phase: the Specification stage, starts with prioritizing User Stories to support the definition of the minimum viable product, with the aim of identifying the core features of BigDataOcean. This will ensure that the resulting platform that is developed is what the clients want and need. This phase therefore includes the specification of the Proof-of-Concept, as well as the architecture. It is therefore possible to identify any discrepancies between the existing requirements and the planned ideas or concepts, and amending them by looping back to Elicitation and Analysis requirements phases.

The fifth and final phase of the implemented methodology has the aim of validating implementations. In this phase, we therefore check that the resulting platform accurately responds to the stakeholder’s needs. Through a technical evaluation we will also verify if the implementations meet the functional and non-functional requirements defined in the previous phases.

A. Big Data Management in BigDataOcean

The BigDataOcean platform enables the ingestion of data in raw format in the data lake, as well as the execution of data management tasks on-demand. Data management tasks include semantic enrichment, and data curation and integration. The BigDataOcean query processing engine executes queries against a data lake and fires the data management tasks required to ensure that query answers are semantically enriched with the BigDataOcean vocabulary. Fig. 7 illustrates a user story implemented at the data management level in the BigDataOcean platform. Users request to know the tank vessels that have arrived in a given port, e.g., the port of Piraeus, in a time period. A SPARQL query represents this request, and the BigDataOcean query processing engine returns the required information containing the specific vessels that arrived in the port of Piraeus. Semantic enrichment of the query answer is performed during query processing. The BigDataOcean query processing engine ensures that only the data that is required to evaluate the query is semantified with the BigDataOcean vocabulary, without introducing a big overhead on execution time. On-demand semantic enrichment avoids that large volumes of data irrelevant for a user story are semantically described. Thus, data storage is utilized efficiently.

B. Big Data Analytics in BigDataOcean

The BigDataOcean platform offers different big data analytics and visualization methods; however, it also allows for the integration of new big data methods. Data required during data analytics and visualization is collected from the data lake using the BigDataOcean query engine; thus, semantic enrichment of the required data is conducted on-demand. Fig. 8 presents user...
In isolated repositories, with diverse data schema and query tasks to the data analysis phase. Thus, the introduction of Data Lakes addresses, in the first place, the problem of disconnected information silos which require data cleansing, aggregation, and transformation in advance, data lakes provide a central repository for raw data that is made available to the user immediately and defer any aggregation or transformation tasks to the data analysis phase. The BigDataOcean query processing engine ensures that on-the-fly semantic enrichment and linking does not affect execution time, while only the data required to answer the query is semantically enriched.

Fig. 7: Example of Query Processing in BigDataOcean. A sample SPARQL query to select the tanker vessels that have arrived into the port of Piraeus this month. The BigDataOcean vocabulary (bdo) is used to semantically enrich and link data about vessels, ports, and vessel routes on-demand. Data is kept in the data lake in raw format and semantified on-the-fly during query execution. The BigDataOcean query processing engine ensures that on-the-fly semantic enrichment and linking does not affect execution time, while only the data required to answer the query is semantically enriched.

VI. RELATED WORK

Unlike data warehouses that provide a common schema and require data cleansing, aggregation, and transformation in advance, data lakes provide a central repository for raw data that is made available to the user immediately and defer any aggregation or transformation tasks to the data analysis phase. The introduction of Data Lakes addresses, in the first place, the problem of disconnected information silos which is the result of non-integrated heterogeneous data sources in isolated repositories, with diverse data schema and query languages. Data lakes eventually guarantee a common access interface to up-to-date and accurate data for any processing and analysis tasks without conveying the development costs of data maintenance, pre-processing, and transformations.

To tackle the data integration problem of heterogeneous data, a few Data Lake systems have been proposed, mainly with focus on data ingestion and metadata extraction and management. For instance, GEMMS (Generic and Extensible Metadata Management System) [3] for data lakes extracts metadata from heterogeneous sources, stores it in an extensible metamodel, and enriches it with semantic annotations in order to provide basic querying support. A few other approaches of personal data lakes [4] keep data from various data sources in raw format in the data lake after serializing them in a common data format. Also, commercial solutions like Microsoft Azure Data Lake [20] are available. However, none of these Data Lake systems consider the semantics of the underlying data or offer limited query processing capabilities over heterogeneous data of different data formats.

The Semantic Web community has invested significant efforts in integrating heterogeneous datasets from various domains by lifting existing tabular data into semantic data and interlinking these datasets with existing LOD datasets [5]. Some well-known examples include the Bio2RDF project for integrating knowledge from various bioinformatics databases and publishing it as RDF documents [6] and the Linked Government Data approach for converting raw government data into high-quality Linked Data [7]. However, the generation of these datasets is related to costly and error-prone transformations based on numerous assumptions [8], e.g., stable schema and data model, data consistency, and semantic annotations. Given the huge amounts of maritime data generated everyday and the different formats at which the data is available, such transformations are not efficient. In our Big Maritime Data
Architecture, maritime data is kept in raw format and allows data integration and query execution on-the-fly.

VII. CONCLUSIONS AND VISION

Various challenges are currently affecting the development of big maritime data applications, depriving stakeholders the ability to exploit the full potential of this data ecosystem. From a technical perspective, these challenges are mainly related to the big data nature of the data sources and their high level of heterogeneity. In this paper, we have described these challenges and presented the particular requirements for applications building on top of an interoperable and integrated big maritime data layer. Thereafter, an architecture for data management of maritime data is described, together with examples of its application in real scenarios derived from the BigDataOcean European project. Relevant data value chains are also detailed in the paper. Moreover, examples of query processing and data analytics applications on top of the presented architecture are illustrated.

In the future, we plan to extensively evaluate the proposed architecture in different concrete scenarios, in alignment with the four BigDataOcean pilots described in this paper.

ACKNOWLEDGMENT

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 732310.

REFERENCES

[1] M. Chen, S. Mao, and Y. Liu, “Big data: A survey,” MONET, vol. 19, no. 2, pp. 171–209, 2014.
[2] S. Kazemi, S. Abghari, N. Lavesson, H. Johnson, and P. Ryman, “Open data for anomaly detection in maritime surveillance,” Expert Systems with Applications, vol. 40, no. 14, pp. 5719–5729, 2013.
[3] C. Quix, R. Hai, and I. Vatov, “GEMMS: A generic and extensible metadata management system for data lakes,” in 28th International Conference on Advanced Information Systems Engineering (CAiSE 2016), 2016, pp. 129–136.
[4] R. Hai, S. Geisler, and C. Quix, “Constance: An intelligent data lake system,” in International Conference on Management of Data, SIGMOD, 2016, pp. 2097–2100.
[5] F. Scharffe, G. Atemezing, R. Troncy, F. Gandon, S. Villata, B. Bucher, F. Hamdi, L. Bihanic, G. Képéklian, F. Cotton, J. Euzenat, Z. Fan, P.-Y. Vandenbussche, and B. Vatant, “Enabling linked data publication with the Datalift platform,” in AAAI 2012, 26th Conference on Artificial Intelligence, W10:Semantic Cities, Toronto, Canada, 07 2012.
[6] F. Belleau, M.-A. Nolin, N. Tourigny, P. Rigault, and J. Morissette, “Bio2RDF: Towards a Mashup to Build Bioinformatics Knowledge Systems,” Journal of Biomedical Informatics, vol. 41, no. 5, pp. 706–716, Oct. 2008.
[7] F. Maali, R. Cyganiak, and V. Peristeras, “A publishing pipeline for Linked Government Data,” in 9th Extended Semantic Web Conference (ESWC 2012), May 2012.
[8] P. N. Mendes, H. Mühleisen, and C. Bizer, “Sieve: linked data quality assessment and fusion,” in Proceedings of the 2012 Joint EDBT/ICDT Workshops, 2012, pp. 116–123.