Automated Port Traffic Statistics: From Raw Data to Visualisation

Luca Cazzanti, Antonio Davoli, Leonardo M. Millefiori
NATO STO Centre for Maritime Research and Experimentation (CMRE)
La Spezia, Italy
Email: {luca.cazzanti,antonio.davoli,leonardo.millefiori}@cmre.nato.int

***Abstract*—We describe how we leveraged best practices in big data processing pipeline design and visual analytics to prototype the Maritime Patterns-of-Life Information Service (MPoLIS), an information product currently under development at the NATO Centre for Maritime Research and Experimentation (CMRE). MPoLIS supports the maritime industry, governments, and international organizations with visual analytics on vessel traffic in seaports. It addresses three main requirements: a) storing and processing large amounts of data; b) on-demand availability of statistical summaries of vessel traffic in ports; c) intuitive and interactive interface for subject matter experts (SMEs) in the maritime domain. MPoLIS has contributed to building a data-driven, self-service analytics culture within NATO and has been sanctioned for use in support of maritime situational awareness (MSA) in ongoing NATO operations.**

I. Background

Readers outside of the maritime shipping industry often do not realize the fundamental role of maritime trade in the global economy: 80% of global trade is by sea, far and away the most dominant method for transporting goods globally. Given its importance, maritime commerce stakeholders are diverse, and include national government, international organizations, policy advisors, security and safety agencies, vessel operators, port authorities, and trade analysts just to name a few.

In the past, a stakeholder seeking to understand maritime patterns could rely on few sources of information, which limited the achievable analyses, but also kept the analyst's daily workflow to manageable levels. Today, this is not the case any more. Big maritime data are now the norm, generated from the Automatic Identification System [1], radar, port visit and vessel registries, and ever-more connected databases and more accessible repositories. We argue that the maritime domain is undergoing a transformation akin to the nascent Internet of Things.

The abundance and availability of maritime data afford the promise of more sophisticated analyses and more timely, actionable insights than previously possible. However, huge amounts of data overwhelm the end user and challenge the established analysis workflows. To make sense of, and gain advantage from big maritime data, stakeholders must rely on summary statistics and representative Patterns of Life (PoLs) to understand the organic behavior of maritime traffic, infer general trends, and extract the key indicators related to trade, safety, regulatory compliance, and day-to-day vessel and port operations. Examples of such indicators are cargo throughput, fishing activity, types of goods transported, piracy and suspicious activity risk assessments, and connectivity measures between ports at local, regional, and global scales.

Researchers from the areas of computer science and analytics have begun addressing the need to extract key indicators from big maritime data. Their work in this area, which we call *computational maritime situational awareness* (MSA) is composed of two main threads. One seeks to develop machine learning algorithms that automatically extract PoLs, predict vessel locations, and detect anomalies [2]. The other focuses on scaling the algorithms to big data regimes and addresses the storage and processing challenges posed by large volumes of data by adopting industry-standard tools like Hadoop, MapReduce, and NoSQL [3], [4].

This work falls in this latter thread of computational MSA, but additionally addresses the problem of delivering the extracted maritime indicators to the end users through an interface that enables interactive exploration of the results. In particular, this work is motivated by the need to simplify the daily workflow of maritime analysts and subject matter experts (SMEs) who are tasked with understanding the PoLs of port traffic and with reporting their findings to higher levels of management. More generally, this work demonstrates to NATO end-users how data-driven, self-service analytics can help address real-world, everyday needs of MSA analysts.

II. Requirements and Challenges

The high level requirements are:

- Port traffic statistics - Given one or more countries and a time range, produce:
  - counts of unique vessels that have visited any ports of the given countries during the given time period;
  - counts of unique vessels broken out by five types: cargo, tanker, fishing, passenger, other;
  - counts of unique vessels broken out by flag state;
  - time series of monthly vessel counts, spanning at least 24 months from the present.
- Automation - The statistics should be updated automatically, on a monthly basis, with no human intervention
- Generalization and scalability - The statistics shall be produced for each country and each port in the

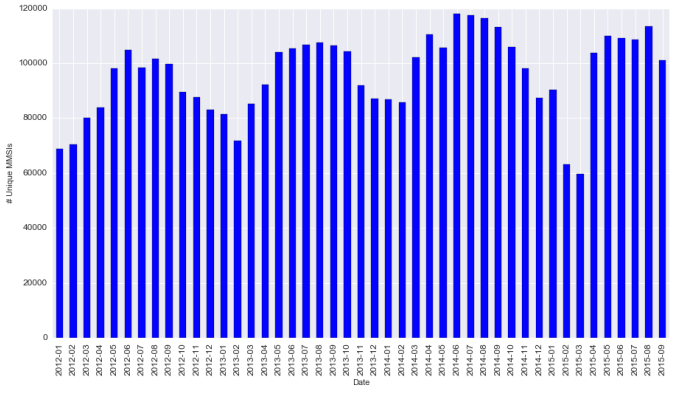

Figure 1. Monthly counts of unique vessels in the historical AIS database maintained at CMRE. In the busiest months, almost $120,000$ vessels are active worldwide.

Mediterranean Sea, and the system shall be scalable to the entire world.

- Information delivery - The statistics should be provided to the end user in a format amenable for inclusion in the user's information sharing processes with minimal formatting or adaptation.
- Data sources - The statistics shall be computed from low-cost, open data sources that reflect the at-sea behavior of vessels.

The above requirements present several challenges. First, the large volume of data forces us to consider computational efficiency when producing the summary statistics for the ports. Consider the number of ports in the world and the number of vessels active on any day. Thousands of ports exist in the world; for example, the World Port Index [5] is an authoritative, but not exhaustive, list of more than $3,600$ ports worldwide. In addition, the number of vessels operating in any given month worldwide approaches $120,000$, as shown in Fig. 1. Furthermore, each month, these vessels transmit several hundreds of millions AIS messages, a number that recently is approaching 1 billion messages/month [4]. Computing the port statistics requires processing this huge amount of data.

Another challenge is related to the way the results are delivered to the end user. In real life, the workflow of a port traffic analyst is slowed down by trivial, yet laborious manual merging operations for data originating form different sources, the lack of ready-to-use charts, and more generally by an overall lack of integrated processes that negatively impact the analyst's ability to share information. Thus, thought must be put into how the analyst will use the port statistics and into ensuring that the solution will be transparent to the analyst's daily workflow.

Additional challenges generated from budget and resource constraints. It was not possible to purchase new hardware to host the application and personnel were not available to reconfigure existing computing assets. Therefore, the newly-developed port statistics processing pipelines would need to be integrated into legacy systems, yet be designed to transition smoothly to dedicated hardware in the future.

III. Technical Approach

To address the above challenges we framed the problem as that of building an information product based on a modular data analytics pipeline and adopted established technologies and best practices from the field of big data:

- A lambda architecture [6] lens for decomposing the problem - The pipeline was conceived from the beginning as a set of modules, following the lambda architecture principles of scalability, generality, extensibility, minimal maintenance, and debuggability. We decomposed the data processing into a batch layer and a serving layer. The port statistics can be updated on a monthly basis with batch processes for which higher latency is acceptable. Results for the latest calendar month can build on previously-computed statistics and batch views are made available to the end user through a thin serving layer. There is no requirement for real-time port statistics, so the speed layer of lambda architecture is not present in our approach.
- MapReduce and parallel processing - Computing the port statistics amounts to filtering, counting, and aggregating vessels appropriately. For scaling these elementary operations to big data regimes, the MapReduce programming paradigm is a perfect match. Thus, in the batch layer, the map and reduce stages can be carried out as distributed operations on a cluster of computers, or as parallel operations on multiple cores on the same node.
- Python language - For the processing pipelines, we adopted the Python programming language because of its flexibility and broad set of specialized modules. It interfaces seamlessly with our existing maritime databases and can run on multiple platforms. It helped us quickly prototype the processing pipelines and visualize the results in the early stages of the development.
- Hybrid cloud infrastructure - The processing pipelines are deployed on compute nodes on premises at CMRE and interface with the existing, on-premises databases. The summary statistics, however, are pushed to the cloud, where the visual interfaces also live. For the serving layer we chose Tableau Online for the summary statistics and a cloud-based Geoserver for the density maps.
- Use AIS as a data source - CMRE has a large database of historical AIS data, which is ideally suited for studying the historical PoLs and highlighting seasonal effects in port traffic. CMRE also receives AIS data form multiple sources, and port traffic statistics can be updated easily with the latest information. Thus, AIS data were essential already freely available to

the project. Finally, AIS data provides most of the information needed to produce the set of basic port statistics desired by the end users: vessel identity, location, characteristics, etc.

- Modern information visualization techniques - In the early stages of the development, we had adopted the Javascript library D3 to bind the user interface with the port statistics. However, modifying the source code to experiment with new information visualization layouts or to incorporate user feedback was time-consuming and error-prone. To iterate more quickly on the user interface design, and to provide an interactive dashboard to the end users, we adopted Tableau, a commercial visual analytics software.

IV. Implementation

Fig. 2 shows the overall structure of the processing pipeline and its component modules. In the Map stage, AIS messages from the CMRE database are pre-filtered and only the vessels that have visited any of the ports under consideration are kept. The dynamic and static vessel information, which are transmitted separately in the AIS protocol, are joined together to obtain the complete vessel information and tagged as "hasVisitedPort". A vessel is tagged as "hasVisitedPort" if it travels at a speed less than 5 knots and enters a 2 km radius around a port's reference coordinates. The coordinates are taken form the World Port Index and are stored in a k-d tree for efficient nearest-neighbor calculation. The choice of the World Port Index, and the speed and radius parameter values reflect commonly-adopted choices by maritime analysts, but are configurable and not mandatory.

The results of the Map stage are saved in intermediate files. In the Reduce stage, the intermediate result files are parsed to produce the batch views, that is final vessel port visit statistics and the density maps. These results are served to the end users through a serving layer, for which we chose a cloud-based browser interface. The candidate dashboards were uploaded to Tableau Online so that end users could provide feedback early in the design process. This choice brought closer collaboration with the end user, and enabled rapid response to feature requests and bug fixes. In addition, with Tableau, the serving layer taps directly into the intermediate files produced by the Map stage and the statistics are computed on the fly, visually. In essence, the Reduce stage is almost entirely delegated to the user interface. This allowed us to experiment quickly with various types of summary aggregation and vessel groupings, for example counting unique vessels vs. counting number of vessel trips. Accomplishing these types of exploratory analyses with D3 is difficult, because D3 is better suited for well-specified visual presentation than for exploratory visual analytics.

Fig. 3 shows an example of one of the dashboards through which end users access the ports statistics. On the left side of the dashboard, the monthly vessel counts are visualized through line charts, broken out by port, ship type, and state flag. On the right side, bar charts report the total vessel counts over a selectable range of dates. Users can drill-down to port-specific statistics by selecting one or more ports from the bar chart. In this particular example, the port of La Spezia has been selected, so the ship type and state flag refer only to this chosen port. Users may further restrict their queries by selecting a subset of ship types and vessel flags (example not pictured). Finally, users may also save screenshots of the dashboard to image files, download the relevant data underlying any of the charts and save them to text files.

Users can navigate from the Tableau dashboard to a density map of vessel traffic produced from the underlying AIS positional reports, as in Fig. 4. Different map layers carry information on different vessel types, and the user can select the specific types to focus the analysis. In this way, the summary statistics for the ports of a country can be related to the actual vessel trajectories. The maps are hosted on a separate map server, and updated on the same schedule as the port statistics, currently monthly.

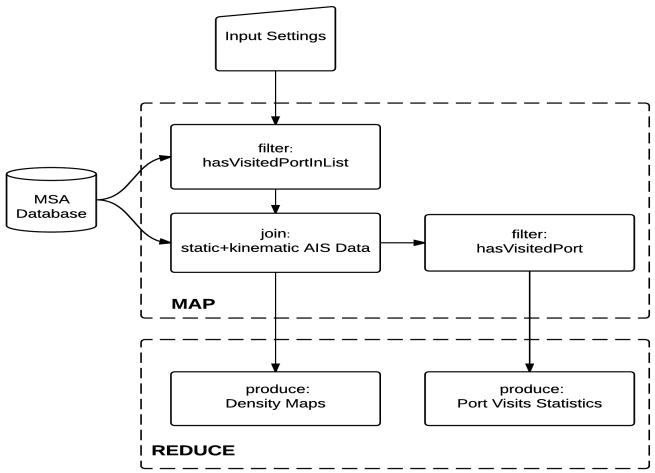

Figure 2. The components of the port statistics batch processing pipeline and their MapReduce roles.

V. Summary

The Maritime Patterns-of-Life Information Service (MPoLIS) provides summary statistics of vessel port visits by leveraging modern big data technologies and practices and the rich information from the Automatic Identification System (AIS). The main utility of MPoLIS is that of improving the workflow of maritime SMEs who must debrief their colleagues on the state of port traffic in given regions of the world. More generally, this work demonstrates how data-driven, self-service analytics can address the challenge of extracting information for the maritime domain. Within NATO, MPoLIS has been sanctioned for use in support

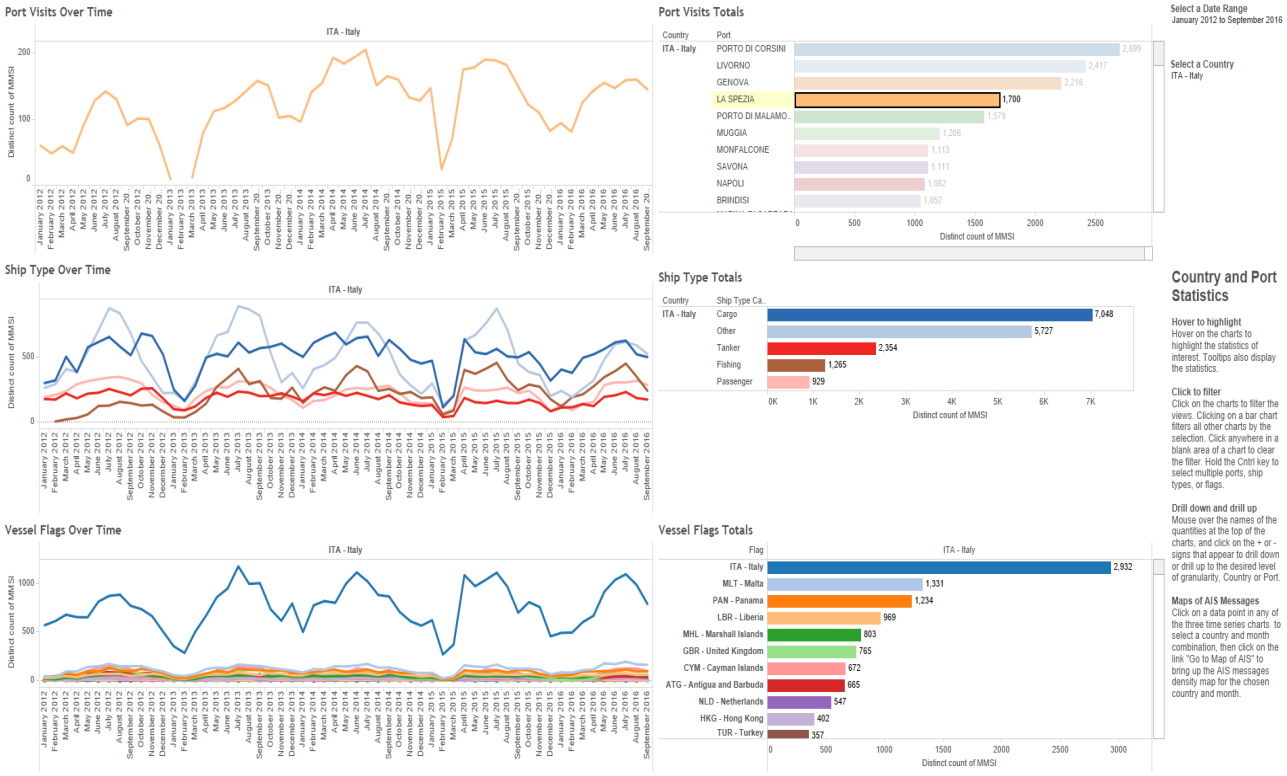

Figure 3. Example Tableau dashboard for the port visit summary statistics. Users can select countries of interest and a time range, and specific ports, to understand the general patterns of vessel traffic in the chosen ports.

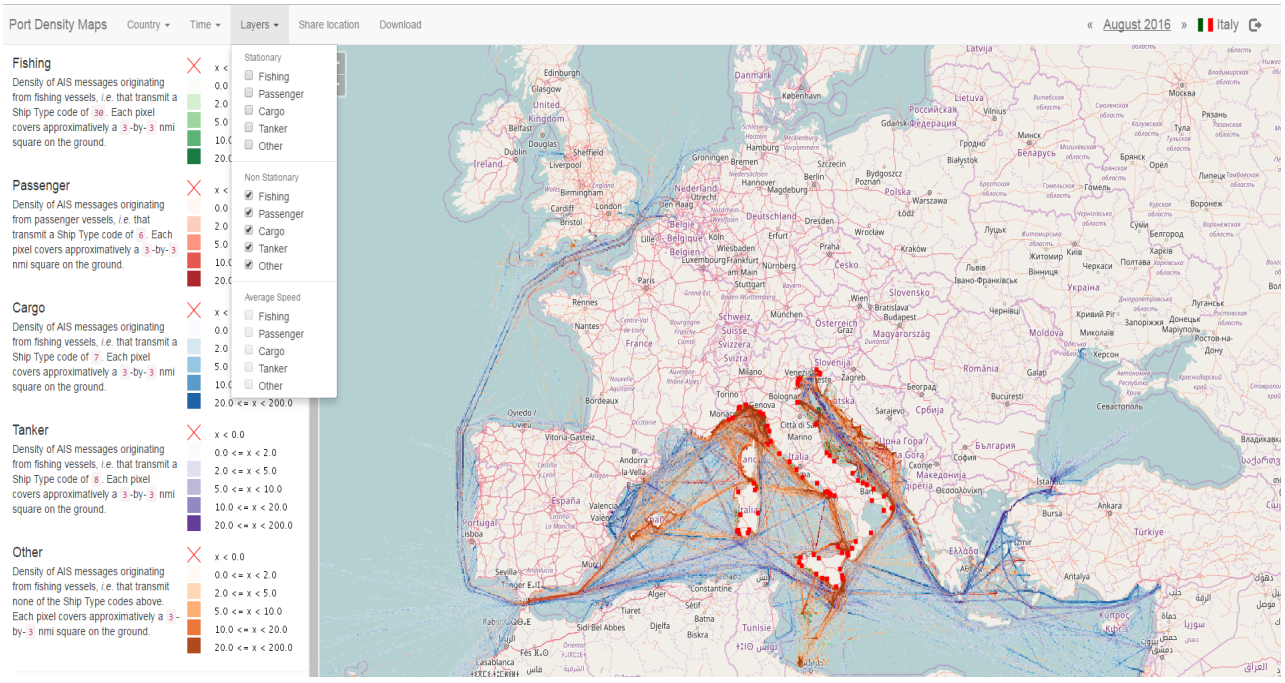

Figure 4. Example density map for the cargo vessels that have visited Italian ports in the Mediterranean. Vessel type can be selected from pull-down menus.

of MSA activities during ongoing maritime patrolling operations. The modular and scalable architecture supporting MPoLIS makes it easily extensible with a richer set of port statistics to provide a more comprehensive picture of port traffic patterns. Finally, the cloud-based Tableau front-end makes MPoLIS immediately deployable to any organization interested in understanding the maritime traffic patterns.

REFERENCES

[1] International Telecommunications Union, "Technical characteristics for an automatic identification system using time division multiple access in the VHF maritime mobile band (Recommendation ITU-R M.1371-4)," 2012.

[2] G. Pallotta, M. Vespe, and K. Bryan, "Vessel pattern knowledge discovery from AIS data: A framework for anomaly detection and route prediction," *Entropy*, vol. 15, no. 6, pp. 2218–2245, 2013. [Online]. Available: http://www.mdpi.com/1099-4300/15/6/2218

[3] B. Liu, E. de Souza, S. Matwin, and M. Sydow, "Knowledge-based clustering of ship trajectories using density-based approach," in *Big Data (Big Data), 2014 IEEE International Conference on*, Oct 2014, pp. 603–608.

[4] L. Cazzanti, L. M. Millefiori, and G. Arcieri, "A document-based data model for large scale computational maritime situational awareness," in *Proc. of IEEE Int. Conf. on Big Data, 2015*. IEEE, October 2015.

[5] NGA, "World port index," National Geospatial-Intelligence Agency, Springfield, Virginia, Tech. Rep. 150, 2015, twenty-fourth Edition.

[6] N. Marz and J. Warren, *Big Data - Principles and best practices of scalable realtime data systems*. Washington, DC: Manning, 2015.