SCIENCE AND TECHNOLOGY ORGANIZATION
CENTRE FOR MARITIME RESEARCH AND EXPERIMENTATION

Reprint Series CMRE-PR-2019-092

A document-based data model for large scale computational maritime situational awareness

Luca Cazzanti, Leonardo M. Millefiori, Gianfranco Arcieri

June 2019

Originally published in:

2015 IEEE International Conference on Big Data, 29 Oct – 01 Nov 2015, Santa Clara, CA, USA, pp. 1350-1356, doi: 10.1109/BigData.2015.7363894

About CMRE

The Centre for Maritime Research and Experimentation (CMRE) is a world-class NATO scientific research and experimentation facility located in La Spezia, Italy.

The CMRE was established by the North Atlantic Council on 1 July 2012 as part of the NATO Science & Technology Organization. The CMRE and its predecessors have served NATO for over 50 years as the SACLANT Anti-Submarine Warfare Centre, SACLANT Undersea Research Centre, NATO Undersea Research Centre (NURC) and now as part of the Science & Technology Organization.

CMRE conducts state-of-the-art scientific research and experimentation ranging from concept development to prototype demonstration in an operational environment and has produced leaders in ocean science, modelling and simulation, acoustics and other disciplines, as well as producing critical results and understanding that have been built into the operational concepts of NATO and the nations.

CMRE conducts hands-on scientific and engineering research for the direct benefit of its NATO Customers. It operates two research vessels that enable science and technology solutions to be explored and exploited at sea. The largest of these vessels, the NRV Alliance, is a global class vessel that is acoustically extremely quiet.

CMRE is a leading example of enabling nations to work more effectively and efficiently together by prioritizing national needs, focusing on research and technology challenges, both in and out of the maritime environment, through the collective Power of its world-class scientists, engineers, and specialized laboratories in collaboration with the many partners in and out of the scientific domain.

NOTE: The CMRE Reprint series reprints papers and articles published by CMRE authors in the open literature as an effort to widely disseminate CMRE products. Users are encouraged to cite the original article where possible.

A Document-based Data Model for Large Scale Computational Maritime Situational Awareness

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

***Abstract*—Computational Maritime Situational Awareness (MSA) supports the maritime industry, governments, and international organizations with machine learning and big data techniques for analyzing vessel traffic data available through the Automatic Identification System (AIS). A critical challenge of scaling computational MSA to big data regimes is integrating the core learning algorithms with big data storage modes and data models. To address this challenge, we report results from our experimentation with MongoDB, a NoSQL document-based database which we test as a supporting platform for computational MSA. We experiment with a document model that avoids database joins when linking position and voyage AIS vessel information and allows tuning the database index and document sizes in response to the AIS data rate. We report results for the AIS data ingested and analyzed daily at the NATO Centre for Maritime Research and Experimentation (CMRE).**

***Keywords*-computational MSA; AIS; MongoDB; NoSQL**

I. Introduction

The Automatic Identification System (AIS) is a collaborative, self-reporting system that allows marine vessels to broadcast their identity, position, and other information to nearby vessels and on-ground base stations. Vessels equipped with AIS transceivers periodically broadcast messages that include the vessel identifying information, characteristics, and destination together with other information coming from onboard equipment, such as current location, speed, and heading [1]. The AIS was originally conceived as a navigational safety to support vessel traffic services in ports and harbours, but soon after its adoption, especially after the International Maritime Organization (IMO) mandated AIS transceivers to be installed onboard a significant number of commercial vessels, [2] AIS began being used also to achieve broader Maritime Situational Awareness (MSA) [1], which is the understanding of the factors that impact the economy, environment, security, and safety of the maritime domain.

Machine learning and data mining researchers have been developing automated maritime traffic analysis techniques based on AIS to support the broader MSA stakeholder community, which consists of civilian organizations tasked with safety, search and rescue, and environmental monitoring of the maritime domain; national and international military organizations concerned with defence and security of maritime infrastructure; port authorities and maritime traffic management agencies; maritime shipping, logistics and insurance companies; financial analysts and commodity traders. The literature on machine learning applications to MSA — which we call *computational MSA* — reports successful case studies of using AIS data to discover and characterize maritime traffic patterns, predict vessel routes, and detect anomalies [2]–[5]. Today, computational MSA has become a necessary component of MSA.

At the same time, maritime traffic and global compliance with the international AIS requirements have steadily increased, and the worldwide network of AIS base stations has grown, producing larger and larger volumes of AIS data [6]. The emergent adoption of satellite-based AIS sensors and the introduction of class B transceivers is expected to accelerate further the trend of AIS toward big data [7]. For example, the NATO Centre for Maritime Research and Experimentation (CMRE) continuously receives, stores, and analyses quasi-real-time streams of the global AIS message traffic from multiple aggregation services, and additionally makes available and processes a real-time stream of AIS messages received by its own local AIS base station. Every month, this amounts to approximately 800 million AIS messages from aggregation sources and over 4 million messages from the local feed, produced by over $100,000$ unique vessels. Fig. 1 shows a density map of the worldwide maritime traffic: producing this image required processing approximately 700 million AIS messages.

The larger volume and increased variety of AIS data bring scalability, complexity, generalization, and interpretability challenges to computational MSA. However, the literature so far has demonstrated feasibility on small AIS datasets from geographically circumscribed areas, and the adopted algorithms may not naturally scale to big data, generalize to arbitrary locations, or keep pace with the real-time processing requirements of MSA use cases. In particular, the interdependency of the learning algorithms with the

[1] ITU Recommendation 1371-4, "Technical characteristics for an automatic identification system using time-division multiple access in the VHF maritime mobile band," ITU, Tech. Rep. Recommendation, 2001.

[2] International Maritime Organization, "International Convention for the Safety of Life at Sea (SOLAS)."

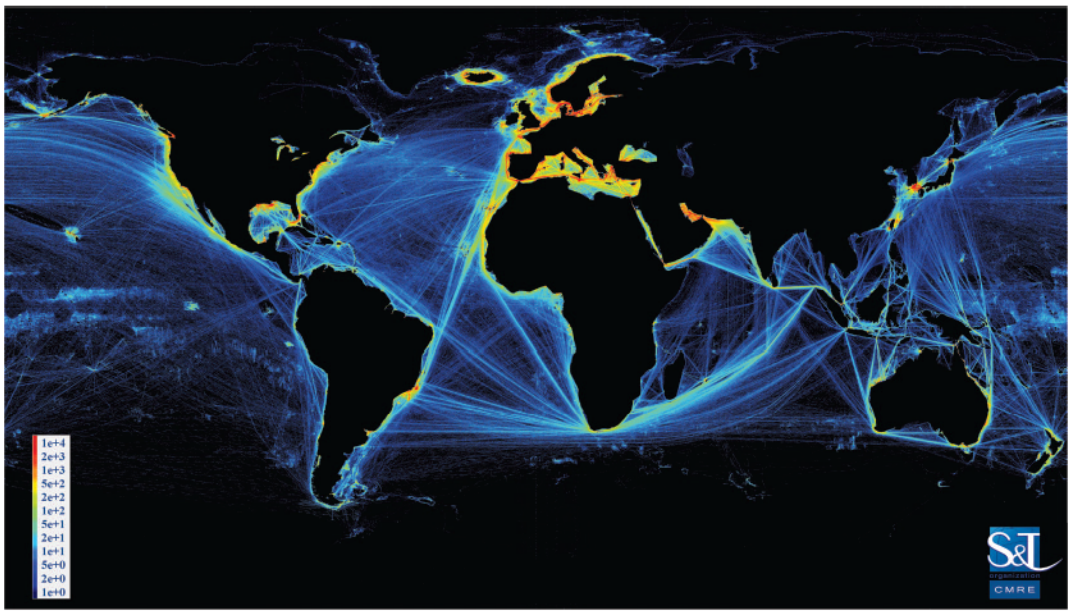

Figure 1. Density map of maritime traffic in April-September 2012, generated from ≈ 2 billion AIS messages collected from multiple networks. Each pixel represents the number of unique vessels that have reported their position within a corresponding 4 nautical mile (one-fifteenth-degree) square cell.

data storage modes and the data models remains largely unexplored. At the same time, achieving MSA from AIS broadcasts critically depends on establishing a clear and timely correspondence between current and historical vessel locations and features, the vessel identifying information, and geographical entities like ports or extended economic zones. Thus, efficient storage and retrieval of AIS data along the time and location dimensions are key enablers for the emergent second wave of computational MSA.

In this paper, we address storage and retrieval strategies for AIS data in support of computational MSA using MongoDB, a widely adopted document-based NoSQL database that pledges horizontal scalability, accommodation of a high insertion rate through sharding and advanced geo-spatial capabilities. First we describe a document-based data model for AIS data that embeds the vessel static-and-voyage information within a document containing a vessel's AIS position report. Then we discuss how the MongoDB database index and the document sizes may be flexibly tuned to the rate of arrival of ASI messages or to the processing needs of downstream computational MSA tasks. Finally, we demonstrate two geospatial queries based on the proposed document data model that represent typical analyses in support of more complex computational MSA. This paper will interest computational MSA practitioners who are interested in exploiting big data technologies in particular NoSQL and document-based data models, and are looking for some guidance; big data architects and data scientists who are interested in building software platforms to serve the burgeoning field of maritime sensor data analytics, in particular vessel traffic AIS data.

II. Background and Related Work

A. Machine learning for maritime traffic analysis

Existing machine learning approaches to vessel traffic analysis can be divided in two categories: point-based and trajectory-based. Point-based approaches treat each AIS position message as a point on a geographical grid. Within each cell, statistical models of the relevant AIS message dimensions — number of messages, number of unique vessels, vessel velocity and direction, etc. — can be estimated and used to detect anomalies, predict vessel locations, build maritime traffic density maps, or characterize AIS sensor performance [4], [8]–[12]. Because the messages in each grid cell are used independently of the ones from other cells to estimate the statistical and predictive models, the underlying binning, counting, and averaging operations, are simpleer, making point-based approaches good candidates for the MapReduce paradigm [13] and for distributed storage modes. However the published literature on integrating point-based approaches with these big data technologies remains scarce.

Trajectory-based approaches relax the independence assumptions and the need for geographic grids and focus on estimating trajectories from the spatio-temporal distribution of each vessel's AIS message stream. The estimated trajectories can then be clustered according to various measures of trajectory similarity, or used as prior knowledge for anomaly detection. Some [14] adopt methods from geometry to estimate the trajectories, but the great majority of the published literature [2], [5] takes a hybrid approach instead, based on the spatial clustering algorithm DBSCAN [15]. The vessel

locations, augmented with speed and direction information, are first clustered, and the resulting small spatial clusters are connected together to form coherent trajectories of groups of vessels [2], [3], [5], [7].

Compared to point-based approaches, forming trajectories requires more complex mathematical operations, careful bookkeeping of the semantic objects — e.g. trajectories, AIS messages, vessel identity, spatial clusters — and tighter integration with storage models. This makes decomposing the underlying operations into MapReduce primitives more challenging [16] and demands careful consideration of how the trajectory learning algorithms should interact with big data distributed storage technologies.

B. Big data techniques for AIS and other sensors

Wijaya and Nakamura [17] demonstrate how HBase can support ship route prediction from AIS data. They predict a vessel's future location from the historical locations of its k-nearest neighbors (k-NNs). The neighbors are vessels of the same ship type, navigational status, and draught, that have previously visited the vessel of interest's current location. Fast access to the k-NNs is achieved by carefully designing the database keys, which exploits HBase's low-latency, random access characteristics. The row key is composed from the ship type, navigational status and draught fields in the AIS messages and the column key is the geohashed latitude and longitude. The algorithm runs on a Hadoop cluster of 11 machines.

Van der Veen et al. [18] assess the performance of SQL and noSQL databases for storing sensor data on physical servers and on virtual machines. They evaluate PostgreSQL, Cassandra, and MongoDB for single- and multiple-client use cases, with and without database indexes. They concluded that for large datasets Cassandra performed best, and MongoDB performed better for medium datasets. PostgreSQL was a better choice when the power of a full-featured query language is needed. That work considered a basic sensor data model (id, timestamp, value) and single-node database installations, which do not fully exercise Cassandra and MongoDB in their ability to horizontally scale for semi-structured data.

Wang at al. [19] demonstrate how to carry out a maritime anomaly detection task on a Hadoop cluster with MapReduce. They take a two-stage approach. First, AIS data are spatially clustered with a variant of DBSCAN, and manually augmented with contextual information from human experts. Then, a parallel meta-learning algorithm implemented with MapReduce detects the anomalies. The detector accuracy and execution time improve linearly with the number of nodes in the cluster, which is consistent with Hadoop's horizontal scalability characteristic.

Implementations of DBSCAN exist that partly address its scalability and parallelization limitations [20], [21]. More recently, He et al. [16] proposed a 4-stage MapReduce implementation of DBSCAN and demonstrated its performance on non-maritime GPS data.

C. Relational and document-based models for AIS

The ITU 1371-4 standard defines 64 different types of AIS messages that can be broadcast by AIS transceivers. In this work we focus on the 6 most relevant ones for MSA, which account for approximately 90% of AIS typical scenarios [22]. Types 1, 2, 3, 18, and 19 are position reports, which include latitude, longitude, speed-over-ground (SOG), course-over-ground (COG), and other fields related to ship movement; type 5 messages contain static-and-voyage information, which includes the International Maritime Organization(IMO) identifier, radio call sign, name, ship dimensions, ship and cargo types. In all messages, each vessel is identified by its Marine Mobile Service Identifier (MMSI). The AIS communication protocol is asynchronous and prescribes that different types of messages be transmitted with different frequencies: static information (type 5 messages) every 6 minutes; position information every 2 seconds to 3 minutes, depending on the speed, location, and navigational status of the vessel.

In a relational database [6], each AIS message is stored as a row in a table whose columns are the message information fields. Standard practice is to store the different types of AIS messages, which contain different fields and arrive at different rates, in two separate tables. Typical computational MSA tasks require linking the vessel information from the two tables, which can be accomplished with a relational join operation. Another option could be to store both types of messages in wider tables, where each row is the union of the information fields from both position and static message types. In this way, computational MSA tasks have direct access to a complete set of attributes about each vessel without requiring a join when the algorithms execute. However, position and static messages are in effect joined to form rows in the wider table at the acquisition stage of the AIS data pipeline. The main advantage of relational databases is the broad availability of mature technologies and a strong tradition in the structures query language (SQL).

In a document database, data are stored in documents. Within a document, information is stored as key-value pairs. Conceptually, each key is analogous to a column name in a relational database, and each value is the row element corresponding to that column. In this way, each document encapsulates both the schema and the data in one, self-contained object. A benefit of the document-based data model is flexibility. First, different types of data streams can be encapsulated in different types of documents that need not abide to the same overall schema. Second, the key-value document structures can be altered dynamically and independently. Third, documents can be embedded in

other documents to create heterogenous information chains that capture complex problem-space semantics.

III. A Document-based Data Model for AIS Data

We adopt a document-based data model to ingest and store AIS messages, using MongoDB. To address the computational MSA requirement to link the information in the position messages with the information in the static-and-voyage messages, we embed both types of information about each vessel in a single MongoDB document[3]. The key-value pairs represent the standard fields of AIS position messages[4] e.g. MMSI (`"msi"`), latitude and longitude (`"lat"`, `"lon"`), speed-over-ground (`"sog"`). The static information is embedded in the key-value pair `"sta"`. The value is a sub-document of key-value pairs from the vessel's last static-and-voyage AIS transmission, e.g. radio call sign (`"csn"`), name (`"nam"`), IMO number (`"imo"`), destination (`"dst"`).

Embedding the static information as an additional key-value pair in the vessel's position report requires managing the asynchronous communications and the different intervals in AIS broadcasts at the time of data ingestion: position messages are much more frequent than static-and-voyage messages and this difference in frequency must be managed. For this, we adopted a buffering procedure where incoming static/voyage messages are held in a buffer that keeps every vessel's latest static information, which is overwritten when a new AIS static/voyage message for that vessel arrives. When a position report arrives, it is matched to the static information stored in the buffer using the MMSI as the match key, and a new document is added to the MongoDB database containing the latest position report with the embedded matched static information.

There are drawbacks and advantages to embedding the static information with the position information in one document. A drawback is that the document size increases. Furthermore, the same static-and-voyage information is duplicated in potentially many different documents, because multiple copies of the same AIS static message are embedded in many position messages from the same vessel. Another cost is the increased complexity introduced by the buffering mechanism in the data ingestion pipeline, which amounts to perform a database join (albeit of a single position message with a single static-and-voyage message) in quasi-real time. These increased storage and complexity costs must be weighted with the convenience and flexibility afforded by the embedded document approach. For many computational MSA tasks, having immediate access to all the information about a vessel in a single document can simplify the data management components of the machine learning algorithms, and thus can help the algorithm developers focus on the semantics of their chosen problem.

Many computational MSA tasks require aggregating AIS data over time ranges at different granularity. Daily, weekly, monthly, and seasonal statistical analyses of maritime traffic are very common, as are minute-resolution anomaly detection and route prediction tasks. To address this need, we consider the MongoDB combined position and static-and-voyage documents from a vessel as samples in a time series of documents, where the timestamp is the time of arrival of the vessel's position message[5]. From the document time series, aggregate vessel statistics over arbitrary periods of time can be computed by aggregating the documents at the appropriate time granularity.

Efficiently accessing the documents by time requires a database index based on the document timestamps. One indexing strategy creates an index in MongoDB for each individual document. Another strategy pre-aggregates multiple documents at a given time granularity and indexes the aggregated documents. In fact, different collections of the same documents pre-aggregated at different levels of granularity may coexist in the same database, which in this way can serve the needs of a range of computational MSA tasks.

We have experimented with document pre-aggregation in our MongoDB setup. Table I shows the document and database index sizes for the individual- and aggregated-document strategies, based on 1 day worth of AIS messages received by the CMRE from 3 different sources with different characteristics. Sources S1 and S2 are AIS data aggregators that stream AIS messages from worldwide locations. S1 downsamples the real-time AIS stream but provides broader worldwide coverage. S2 provides the raw AIS stream from a more limited worldwide coverage. S3 comes from CMRE's local AIS base station, which provides a real-time raw data stream of AIS traffic in the Gulf of La Spezia, Italy.

A window of 60 seconds was chosen for the time-aggregated strategy, which results in a fixed index size of 1,440 documents per day, equivalent to approximately 50 KB. The sizes of the minute-aggregated documents vary depending on the data source, and reflect the rate at which AIS messages arrive at the CMRE computing facilities and the area of coverage of the AIS messages. For the single-document strategy, the average document size is approximately 1 KB, but the index size varies depending on the source. There is an order of magnitude difference in the document and index sizes for the individual- and aggregated-document strategies.

There are benefits and costs to each of these strategies.

[3]An example document, expressed in the JavaScript Object Standard format (JSON), can be accessed at https://goo.gl/2jCG6G.

[4]It is standard practice to abbreviate the key names to minimize the document size.

[5]AIS messages contain only partial time information, so it is standard practice to assign each message a full UTC timestamp when it arrives at the computing facilities. The propagation time from the transmitter to the receiving station is typically negligible.

Table I
DOCUMENT AND INDEX SIZES FOR INDIVIDUAL- AND AGGREGATED-DOCUMENT STRATEGIES.

	S1		S2		S3	
	individual	aggregated	individual	aggregated	individual	aggregated
Count	4 703 244	1439	2 040 082	1439	632 360	1439
Average document size	0.89 KiB	3.99 MiB	0.86 KiB	7.94 MiB	0.98 KiB	0.46 MiB
Index size	130.99 MiB	0.05 MiB	63.13 MiB	0.05 MiB	17.63 MiB	0.05 MiB

In the single-document strategy, many small documents are stored, giving access to individual AIS messages at a finer granularity, but the database index size is necessarily larger. In the aggregated-document strategy, each document is larger and gives access to pre-aggregated AIS messages at a coarser time resolution, and the index is smaller. Furthermore, the number of pre-aggregated documents is fixed, because the number of seconds, minutes, hours, days in a given time range does not depend on the number of AIS messages received. Document and index size are important competing parameters to consider when configuring the database. In particular, if the index grows larger than the available memory, database performance will degrade. Furthermore, MongoDB currently has a hard limit of 16 MB for each document.

Notwithstanding the database configuration issues, a more fundamental question is: When is pre-aggregation useful for computational MSA? It depends on the particular task. On one hand, aggregating documents at a fixed time resolution provides faster access to sets of documents within specified time ranges, which can lower the complexity of computational MSA tasks that need to access batches on AIS messages grouped by time. On the other hand, aggregation does not provide immediate access to the individual attributes in the aggregated documents, which must be unpacked and made available individually to the algorithms. This increases the complexity of the learning algorithms that need to access the data. From these considerations, we hypothesize that no aggregation or finer aggregation levels are better suited for near-real-time processing, like anomaly detection, where the full vessel details must be readily and simply available to the computational MSA processes. Coarser aggregations are better suited for computational MSA tasks based on batch processing , such as characterizing the historical patterns of life with spatial clustering, or generating a vessel traffic density map. For these tasks, higher latency is acceptable and indeed helps manage the increased complexity.

IV. GEOSPATIAL ANALYSES

We illustrate how to interact with document-based AIS data through examples of typical geospatial analyses. We leverage the geospatial capabilities built in MongoDB and the embedded-document model that combines position and static AIS messages, and demonstrate two typical queries in support of more complex computational MSA. For these demonstrations, we installed MongoDB on a computer equipped with a 16-core, 2.4GHz Intel Xeon CPU, 32 GB RAM and a 480 GB solid state drive. We loaded MongoDB with 39 days of data spanning 1 May 2015 to 9 June 2015, totaling approximately 241 million AIS messages. These data occupied 201 GB of disk space, and the database index, built using a single. document strategy, occupied 39 GB.

A. Vessel Traffic Characterization in a Non-convex Region

Fig. 2 shows the vessel tracks in the Strait of Gibraltar, measured from AIS transmissions over a 5-day period starting 3 May 2015. The outer, dotted-line bounding box and the inner solid-line one define a non-convex region of interest. Within the inner bounding box, the pixels forming the vessel tracks are color-coded by the speed reported by the vessels at the pixel coordinates. Note how the speeds decrease as the tracks approach the Gibraltar choke point, where more tracks are blue and green (slower speeds). Note also that some tracks away from the choke point are blue (slowest speed) and describe self-contained patterns: these may be fishing vessels or loitering vessels, or assigned waiting areas or off-shore platforms. An interesting track is the blue one at (7.2W, 35N): it is a slow vessel in the open seas, where typical speeds are higher. Fig. 3 shows the distribution of average vessel speeds within the inner bounding box, broken out by ship class. Note that calculating this result required matching AIS static information (ship class) with AIS position information (latitude, longitude, speed-over-ground), which was easily achieved by the embedded document model.

The corresponding query[6] to extract the documents within the inner polygon leverages MongoDB's built-in `$geowithin` construct to filter the documents by the location (`$loc`) key-value pair with respect to the user-defined `$geometry` polygonal bounding box `poly_in`. The query also filters the documents along the time dimension, by checking that the document timestamp (`$tst`) key-value falls within user-defined starting and ending dates.

[6]A sample query is available at https://goo.gl/VFPZ2e.

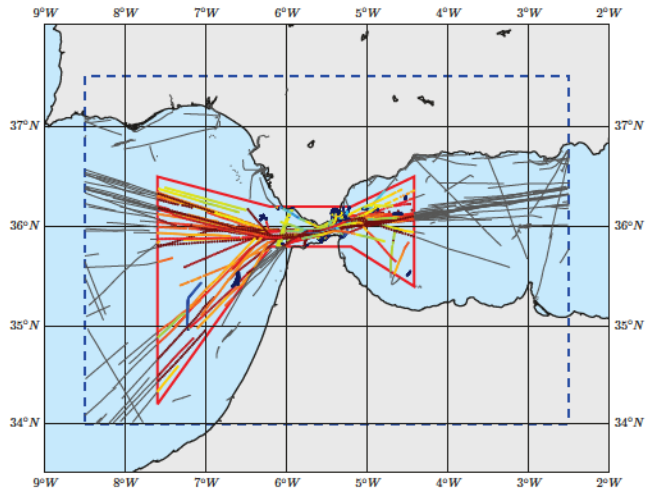

Figure 2. The vessel tracks color-coded by the speed-over-ground AIS attribute, for the Strait of Gibraltar, within a non-convex region of interest.

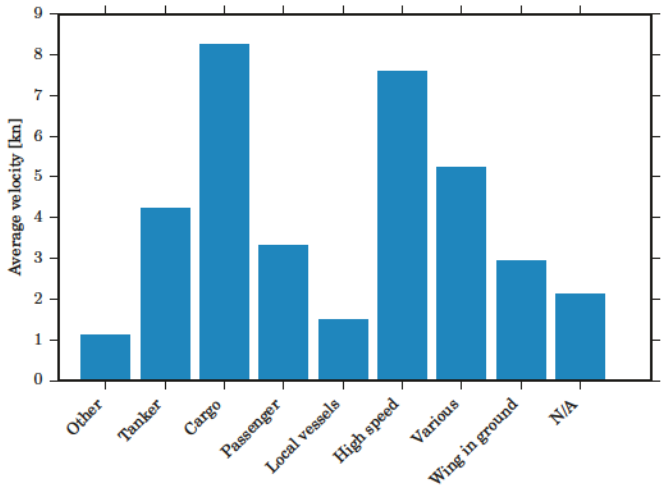

Figure 3. The distribution of average vessel speeds across various types of vessels in the Strait of Gibraltar.

B. Vessel Rendezvous Detector

A vessel rendezvous occurs when two or more vessels transit within a certain distance of each other. This could be the result of normal traffic patterns in a given region, but it could also be an indication of illicit activity, such as transferring goods from one vessel to another without authorization, or transferring illegal cargo. Fig. 4 shows the result of a rendezvous detector run on 10 days of AIS data. The green track refers to a vessel of interest; the other tracks refer to vessels that triggered a rendezvous detection with the vessel of interest. The red circles indicate the actual rendezvous locations, where a rendezvous is detected if a vessel transited within 2 nautical miles but no less than 200 meters of the vessel of interest, and the corresponding AIS

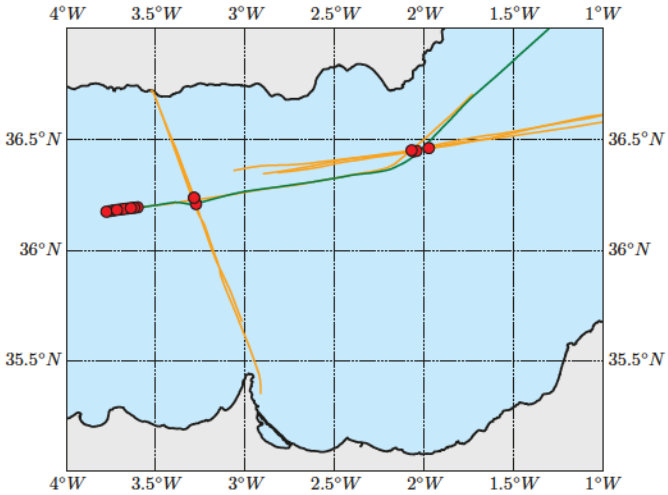

Figure 4. The rendezvous detections.

position messages were timestamped within 400 seconds of each other. The corresponding query[7] leverages the `$near` construct in MongoDB to determine geographical distance.

V. Summary and Future Work

We have experimented with MongoDB, a document-based NoSQL database, as a tool to support the data storage and algorithm development needs of computational MSA. We proposed a document-based data model for AIS data that embeds both position and static-and-voyage types of AIS messages from a vessel in a single document, and described how these embedded documents could be further aggregated by time and stored together at different granularities. We demonstrated two typical queries of AIS data stored in documents on a test MongoDB environment. The main benefit of a document-based data model for computational MSA is flexibility: different types of AIS messages can be stored together or separately, and messages can be stored individually or aggregated by time. This flexibility can help the computational MSA community, which is beginning to study how maritime traffic analysis learning algorithms should interact with big data storage technologies in order to meet the challenges brought by the massive amounts of AIS data.

This work is preliminary, and it did not aim to comprehensively solve the problem of integrating big data technologies with core computational MSA learning algorithms. Possible future work is to evaluate a spectrum of NoSQL tools for computational MSA, in particular how their assumed data models may benefit various computational MSA tasks. A future direction of particular interest is to address how spatial clustering algorithms may make use of document-based storage models and of modern geospatial query tools

[7]A sample query is available at https://goo.gl/SV6Ra9.

such as the ones we experimented with using MongoDB. This effort could lead to new algorithms that can potentially deeply influence the future of computational MSA.

ACKNOWLEDGMENT

L. Cazzanti, L. M. Millefiori and G. Arcieri would like to thank the NATO Allied Command Transformation (ACT) for supporting the CMRE project Data Knowledge Operational Effectiveness (DKOE).

REFERENCES

[1] B. Tetreault, "Use of the automatic identification system (AIS) for maritime domain awareness (MDA)," in *OCEANS, 2005. Proceedings of MTS/IEEE*, Sept 2005, pp. 1590–1594 Vol. 2.

[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] F. Mazzarella, M. Vespe, D. Damalas, and G. Osio, "Discovering vessel activities at sea using AIS data: Mapping of fishing footprints," in *Information Fusion (FUSION), 2014 17th International Conference on*, July 2014, pp. 1–7.

[4] B. Ristic, "Detecting anomalies from a multitarget tracking output," *Aerospace and Electronic Systems, IEEE Transactions on*, vol. 50, no. 1, pp. 798–803, January 2014.

[5] 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.

[6] G. Cimino, G. Arcieri, S. Horn, and K. Bryan, "Sensor data management to achieve information superiority in maritime situational awareness." CMRE, Tech. Rep., 2014.

[7] N. Le Guillarme and X. Lerouvreur, "Unsupervised extraction of knowledge from s-ais data for maritime situational awareness," in *Information Fusion (FUSION), 2013 16th International Conference on*, July 2013, pp. 2025–2032.

[8] B. Ristic, B. La Scala, M. Morelande, and N. Gordon, "Statistical analysis of motion patterns in AIS data: Anomaly detection and motion prediction," in *Information Fusion, 2008 11th International Conference on*, June 2008, pp. 1–7.

[9] M. Zandipour, B. Rhodes, and N. Bomberger, "Probabilistic prediction of vessel motion at multiple spatial scales for maritime situation awareness," in *Information Fusion, 2008 11th International Conference on*, June 2008, pp. 1–6.

[10] B. Rhodes, N. Bomberger, and M. Zandipour, "Probabilistic associative learning of vessel motion patterns at multiple spatial scales for maritime situation awareness," in *Information Fusion, 2007 10th International Conference on*, July 2007, pp. 1–8.

[11] F. Deng, S. Guo, Y. Deng, H. Chu, Q. Zhu, and F. Sun, "Vessel track information mining using AIS data," in *Multisensor Fusion and Information Integration for Intelligent Systems (MFI), 2014 International Conference on*, Sept 2014, pp. 1–6.

[12] G. Papa, S. Horn, P. Braca, K. Bryan, and G. Romano, "Estimating sensor performance and target population size with multiple sensors," in *Information Fusion (FUSION), 2012 15th International Conference on*, July 2012, pp. 2102–2109.

[13] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," in *Proc. 6th Symposium on Operating Systems Design and Implementation*. San Francisco, CA: USENIX, December 2004.

[14] G. K. D. de Vries and M. van Someren, "Machine learning for vessel trajectories using compression, alignments and domain knowledge," *Expert Systems with Applications*, vol. 39, no. 18, pp. 13 426 – 13 439, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417412007762

[15] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Second International Conference on Knowledge Discovery and Data Mining*, Portland, OR-US, 1996.

[16] Y. He, H. Tan, W. Luo, H. Mao, D. Ma, S. Feng, and J. Fan, "MR-DBSCAN: An efficient parallel density-based clustering algorithm using MapReduce," in *Parallel and Distributed Systems (ICPADS), 2011 IEEE 17th International Conference on*, Dec 2011, pp. 473–480.

[17] W. Wijaya and Y. Nakamura, "Predicting ship behavior navigating through heavily trafficked fairways by analyzing AIS data on Apache HBase," in *Computing and Networking (CANDAR), 2013 First International Symposium on*, Dec 2013, pp. 220–226.

[18] J. van der Veen, B. van der Waaij, and R. Meijer, "Sensor data storage performance: SQL or NoSQL, physical or virtual," in *Cloud Computing (CLOUD), 2012 IEEE 5th International Conference on*, June 2012, pp. 431–438.

[19] X. Wang, X. Liu, B. Liu, E. de Souza, and S. Matwin, "Vessel route anomaly detection with Hadoop MapReduce," in *Big Data (Big Data), 2014 IEEE International Conference on*, Oct 2014, pp. 25–30.

[20] X. Xu, J. J"ager, and H.-P. Kriegel, "A fast parallel clustering algorithm for large spatial databases," *Data Mining and Knowledge Discovery*, vol. 3, pp. 263–290, September 1999.

[21] E. Januzaj, H.-P. Kriegel, and M. Pfeifle, "Scalable density-based distributed clustering," in *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, ser. PKDD '04. New York, NY, USA: Springer-Verlag New York, Inc., 2004, pp. 231–244. [Online]. Available: http://dl.acm.org/citation.cfm?id=1053072.1053095

[22] P. Last, C. Bahlke, M. Hering-Bertram, and L. Linsen, "Comprehensive analysis of automatic identification system (AIS) data in regard to vessel movement prediction," *The Journal of Navigation*, vol. 67, pp. 791–809, 9 2014.

Document Data Sheet

Security Classification		Project No.
Document Serial No.	**Date of Issue**	**Total Pages**
CMRE-PR-2019-092	June 2019	7 pp.

Author(s)

Luca Cazzanti, Leonardo M. Millefiori, Gianfranco Arcieri

Title

A document-based data model for large scale computational maritime situational awareness

Abstract

Computational Maritime Situational Awareness (MSA) supports the maritime industry, governments, and international organizations with machine learning and big data techniques for analyzing vessel traffic data available through the Automatic Identification System (AIS). A critical challenge of scaling computational MSA to big data regimes is integrating the core learning algorithms with big data storage modes and data models. To address this challenge, we report results from our experimentation with MongoDB, a NoSQL document-based database which we test as a supporting platform for computational MSA. We experiment with a document model that avoids database joins when linking position and voyage AIS vessel information and allows tuning the database index and document sizes in response to the AIS data rate. We report results for the AIS data ingested and analyzed daily at the NATO Centre for Maritime Research and Experimentation (CMRE).

Keywords

Computational MSA, AIS, MongoDB, NoSQL

Issuing Organization

NATO Science and Technology Organization
Centre for Maritime Research and Experimentation

Viale San Bartolomeo 400, 19126 La Spezia, Italy

[From N. America:
STO CMRE
Unit 31318, Box 19, APO AE 09613-1318]

Tel: +39 0187 527 361
Fax:+39 0187 527 700

E-mail: library@cmre.nato.int