A Cloud Architecture for big data analytics and real-time anomaly detection in global maritime applications

Mariano Alfonso Biscardi[a], Marcello Cinque[b], Marco Corsi[a], Filippo Daffinà[a], Raffaele Della Corte[b,c], Alfonso Di Martino[c], Claudio Perrotta[c], Salvatore Recano[c], Dino Quattrociocchi[c]
[a]e-Geos S.p.a., via Tiburtina 965, Roma, Italy ; [b]Dept. of Electrical Engineering and Information Technology, Univ. of Naples, via Claudio 21, Naples, Italy; [c]Critiware S.r.l., via Carlo Poerio 89A, Naples, Italy

ABSTRACT

The availability of global and live vessel data streams, improved in time and space resolution, acquired both trough automatic tracking systems (AIS) and satellite target detection, enables a large set of applications, ranging from real-time monitoring of vessel positions to the detection of anomalous behaviours, such as, entering forbidden areas, meeting at sea or sudden change of heading or speed. However, providing both real-time and historical analytics, on significant datasets (e.g., years of data acquisitions) poses severe technical challenges, in terms of performance and scalability. To tackle these issues, in this paper we present a software architecture, underlying a software product named SEonSE, designed to manage large amount of historical vessel data and enabling real-time big data analytics worldwide, with the aim to serve different stakeholders for different purposes. The architecture has been implemented and deployed in the cloud and its functioning has been tested on real live datasets, including global AIS data and satellite vessel detection.

Keywords: Cloud, big data analytics, anomaly detection, maritime, vessel detection

1. INTRODUCTION

The use of the sea as main mode of goods transportation worldwide is progressively increasing over the years. Suffice to think that, according to Eurostat, in 2016 the 80.8% of European exported goods have been transported over the sea, with an increasing trend of about 8% in ten years over other modes of transportation [1]. At the same time, the availability of global vessel identification information coming from "cooperative systems" (e.g. AIS, VMS or LRIT), coupled with vessel information extracted from satellite data (that allow to uncover non-cooperative, or so-called "dark" vessels), is enabling a wide set of applications in the fields of global maritime situational awareness, including the detection of anomalous behaviours [2]. The increase of global sea use also corresponds to an increase in size and density of vessel data available worldwide. For this reason, the use of big data solutions in this domain is gaining momentum [3].

As a matter of fact, the data source that current and future maritime intelligence systems have to handle possesses all the three basic "V"s characterizing a "big data" data set [4], i.e.: (i) *volume*, as gigabytes of AIS and satellite derived vessel data are acquired at global level and on a daily basis; these data need to be properly stored and further analyzed to perform anomaly detection; (ii) *velocity*, as the new data are updated with a pace that can reach a worldwide update at most every 5 minutes; and (iii) *variety*, as multiple data sources are encompassed to build the intelligence required to perform the intended analysis (e.g., to compute an "index risk" associated to each vessel).

In this paper, we present our experience in this direction, presenting the recent evolution of the SEonSE platform towards a cloud-based architecture for big data analytics over worldwide, real-time vessel data. SEonSE (standing for Smart Eyes on SEas) is an e-Geos software product conceived to provide web services, and a related map based web interface, for real-time data and intelligence on vessel information acquired through a variety of sources, such as, the automatic tracking system (AIS), encompassing both collaborative and dark vessels, the latter determined by the on-line analysis of multi-sensor satellite imagery. The first versions of the platform, based on traditional approach and tools (e.g. relational database), were conceived to be used on limited areas of interest over a limited period of time. The need of real-time processing and analysis of worldwide vessel information for global anomaly detection called for a complete re-visitation of the underlying software architecture, towards a cloud-based scalable solution for big data analytics. According to the historical data in our possess, global collaborative vessels data increased of 895% in 5 years, ranging from about 480

GBs in 2013 up to 4.3 TBs in 2018, encompassing about 4 billions vessel data updates per year. On average, about 12 GBs of data are produced everyday, accounting for about 13 millions vessel data updates per day, with frequent updates (in the order of few minutes).

Among the many technical issues arising from the management of such massive amount of real-time data, we specifically focus on the following two: 1) sustained ingestion, i.e., the ability to sustain the pace of updates, being capable to ingest and analyze all new data on time, before the start of the subsequent update; 2) fast response to anomaly queries, i.e., to provide answers to anomaly detection requests, over a given spatial area and in a given time interval, in a reasonable time, without impacting the real-time ingestion; 3) capability to online extract maritime patterns of life by aggregating and fusing all the available information at run-time . To face these issues, the SEonSE cloud architecture is designed to perform scalable on-line data gathering and processing, using in-memory optimizations, and to execute batch pre-processing jobs in the background. This pre-processing approach enables to elaborate and store the outputs of analytics queries in advance, as the ingestion progresses, so to provide almost immediate answers to analysts, if not for the time needed to transfer the result of queries over the network.

In the rest of the paper, we present some details of the cloud-based architecture for big data processing underlying SEonSE (section 2), and provide an initial view of results, focusing especially on the processing time of the global live AIS data stream (section 3). Final remarks and outlook for future work is presented in section 4.

2. SEONSE CLOUD ARCHITECURE

Figure 1 presents the main elements of the SEonSE cloud architecture for big vessel data analytics. In its current implementation, it is based on (and hosted as) Amazon Web Services (AWS) [5], and, as such, it re-uses some key elements, such as S3 for storage, the NAT Gateway for internal re-routing, and the Application Load Balancer.

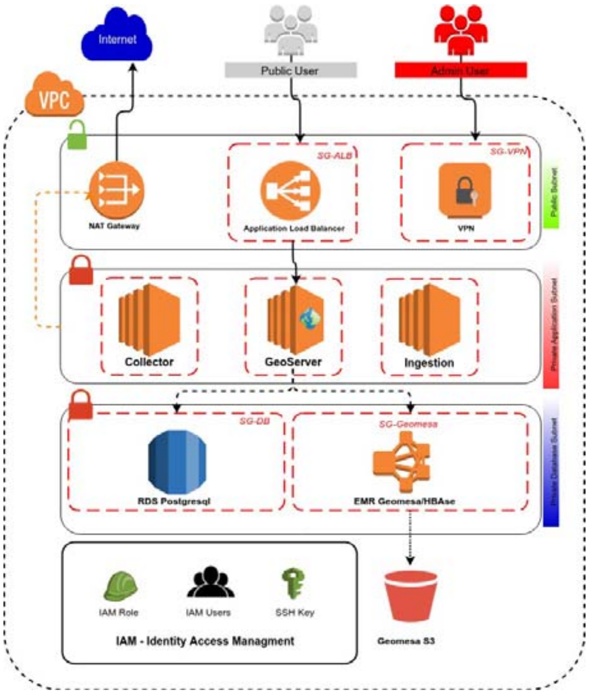


Figure 1. SEonSE Cloud Architecture for big vessel data analytics

The architecture presents 4 main layers, described in the following, starting from the lowest:

- Storage and IAM: at this layer we find the storage of data S3 buckets [5] (used to store raw data and database files, both for Postgres and HBase) and the management of the users along with their roles and privileges to access the platform.

- Private Database Subnet: at this layer we deploy the database components on Elastic Compute Cloud nodes (EC2) [5] in a private subnet, namely a postgresql relational database and a Geomesa/Hbase instance [6] (encompassing a master and more slave nodes). The postgresql relational database, extended with postgis, is used to store configuration metadata and the latest snapshot of worldwide vessel data, particularly useful for the on-line use of the platform. Geomesa, instead, is used to store and retrieve historical vessel data and pre-processed analytics. Geomesa is an open source tool for large-scale geospatial storage, querying and analytics on distributed computing systems. It provides spatio-temporal indexing on top of several nosql database solutions, among which Hbase [7], which we chose due to its inherent scalability, its automatic table sharding capabilities, and the native support of the Hadoop Distributed File System (HDFS). We deployed Geomesa on top of the ElasticMapReduce AWS service to take advance of its built-in scalability support and the native HBase support.
- Private Application Subnet: the private subnet implemented at this layer hosts application level components, among which:
    - The Collector, which gathers vessel data in real-time from external providers (e.g., AIS vendors) and stores them in raw format (also for backup purposes) on S3;
    - The Ingestion, that, starting from the last update provided by the Collector, performs the core operations of the platform, namely, producing the live global vessel snapshot on postgresql, updating the historical information on Geomesa, and computing (and storing on Geomesa) anomaly detection events;
    - GeoServer, a de-facto standard open-source solution to expose collected and processed data (both coming from postgresql or Geomesa) as OGC standard web services, such as WMS and WFS.
- Public Subnet: at this layer we find the components used to expose the managed data to final users, using GeoServer in a replicated fashion (to improve both the performance and the availability) through the Amazon Application Load Balancer service. At this layer, it is also deployed a Virtual Private Network to let admin users access and configure the system.

The architecture share similarities with other initiatives in the maritime domain [8][9], confirming once more the need of the technological shift on the SEonSE platform. The focus of this paper is on the opportunity to perform pre-processing and anomaly computation already during the ingestion, as presented in next subsection, in order to have data about anomaly ready to be queried without further on-line computation.

2.1 The Ingestion and pre-processing chain

Among the architectural components presented, the Ingestion plays a central role. It includes both a processor task to produce the HBase table of historical vessel data and a set of analytic tasks that produce tables of pre-processed data items containing detected anomalies. The implemented chain is exemplified in Figure 2.

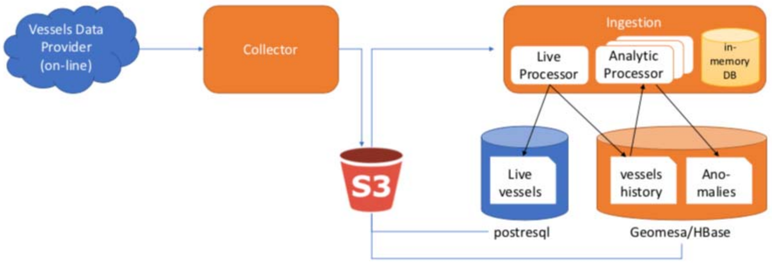


Figure 2. The ingestion chain

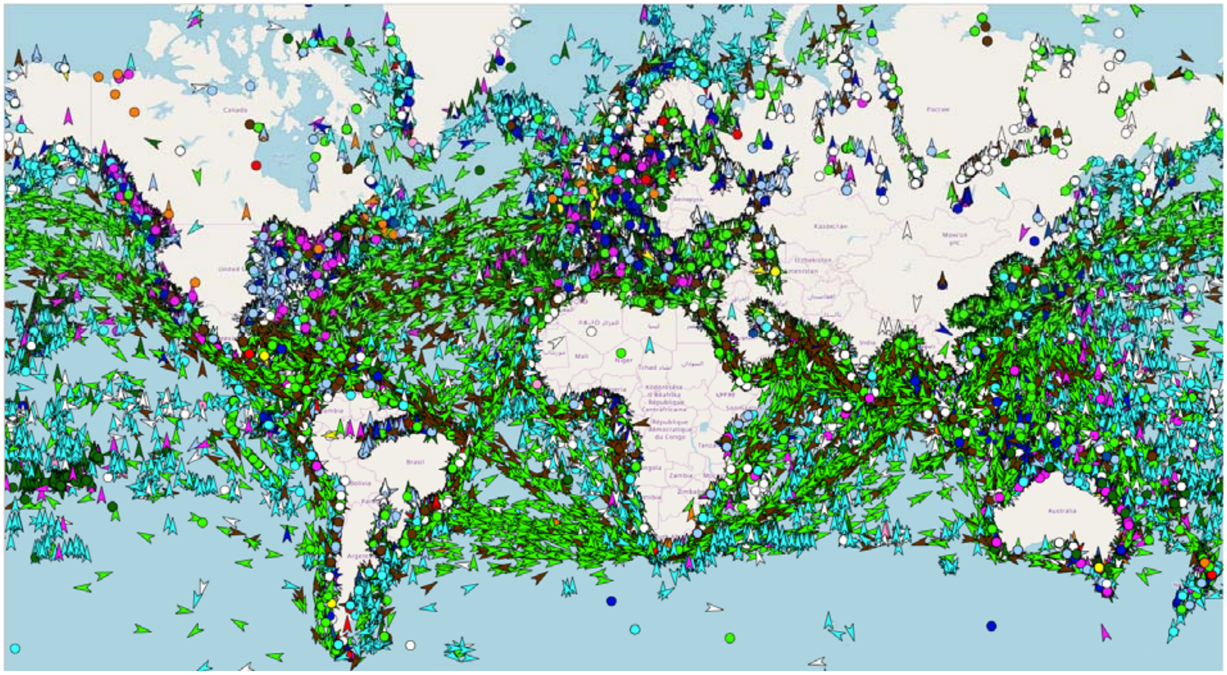


Figure 3. The global AIS vessel live data provided by SEonSE

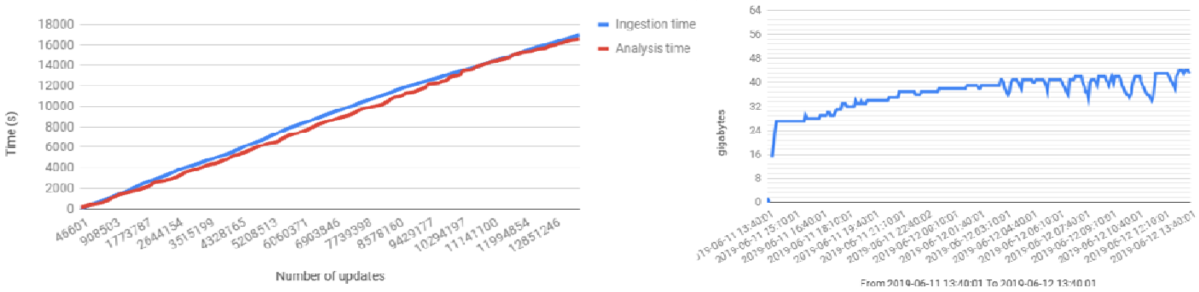
Worldwide vessel data (encompassing about 300.000 vessels, on average) are gathered on-line by the collector every at most 5 minutes, and saved on S3, which acts as a local backup storage and store for further processing, if not covered already by pre-processing. As soon as a new data instance is saved, the ingestion processes it (through the Live Processor, that continuously waits on newly published contents in S3) and (i) produces the Live vessels snapshot in PostgreSQL, containing always the newest real-time information about vessels, and (ii) updates the vessels history with the newly acquired data. The history accumulates the data of vessel during the overall monitoring period and can be used to perform historical queries, in a given time interval, on peculiar vessels or on a given geographical area of interest (AoI).

The vessel history is then processed by Analytic Processors, that implements several anomaly detection algorithms and produce different types of anomaly events in Geomesa/Hbase. Similarly to the history, the anomalies table can be queried over a given event type, time interval and AoI to retrieve current or past information about vessels' anomalous behavior, such as, entering/exiting an AoI, standing in a AoI, performing a change of name or a sudden change of speed or heading, a rendezvous between vessels, etc. Details on algorithms are omitted here due to company restrictions.

The key advance of the solution is that anomalies are pre-processed during the ingestion. Hence, queries on anomalies do not require to actually elaborate the analysis algorithm, but are read operations from Geomesa, taking advance of its optimized indexes. To keep-up with the high pace and volume of acquired data, the processors are equipped with in-memory databases of vessels information and status, in order to quickly perform the required elaboration (without accessing the databases on the permanent storage) and to store only the results (e.g, anomalies found) on HBase, at the end of a transaction.

3. LIVE INGESTION TEST RESULTS

We implemented, configured, and launched the ingestion chain for global, live AIS data on two Amazon EC2 instances with 2 vCPUs and 64GB RAM. A new snapshot is produced by the Collector on S3 every 3 minutes, each containing about 30000 AIS vessels updates. The execution of the whole ingestion and collection chain completes in about 70 seconds, on average, including the reading of the collected snapshot from S3, the reordering of messages, the update of the global AIS vessel live data (shown in Figure 3), the update of the vessel history table on Geomesa/HBase, and the pre-processing and storage of anomalies. Hence, in the current settings, the implemented chain is able to keep-up with the pace of updates, producing meanwhile pre-processed data about anomalies (with insert operations of new data that do not block concurrent read operations) that can be subsequently and efficiently queried from Geomesa, without requiring any further processing.



(a)

(b)

Figure 4. (a) Ingestion and analysis times evolution, and (b) memory usage, over 1 day of collection and analysis

Figure 4.a depicts the evolution of ingestion and analysis times over 1 day of collection. As it can be seen from the figure, times increase linearly with the number of AIS vessel updates, with no observable degradation phenomena over 1 day of operation.

We observe a corresponding increase of the usage of the memory during the process (Figure 4.b), due to our in-memory database, which however stabilizes around 40 GB. Small increases of memory usage are still observable after the stabilization, suggesting us to perform periodic memory maintenance operations (e.g., forcing application reload) once or twice a week. Over 1 day of operation (with data acquired in June 2019), about 13 millions updates are managed, with a total processing time (including both the ingestion and analysis) of about 9.5 hours.

4. CONCLUSIONS

In this paper we presented the architectural evolution of SEonSE, an e-GEOS product for maritime situational awareness and intelligence applications, needed to perform real-time maritime big data ingestion and analytics of vessel information. The proposed solution takes advantage of both state-of-art open-source solutions for the management of large amounts of geospatial information, such as Geomesa/Hbase on top of an HDFS cluster, and in-memory optimizations, to keep-up with the pace of updates. The solution has been implemented and deployed in the cloud, using AWS. Preliminary test results show the practical feasibility of the solution and demonstrate that, with a reasonable amount of computing resources, it is possible to sustain the pace of worldwide live vessel updates, while producing vessel histories and anomalies.

Future work will be devoted to the ingestion and pre-processing of historical data, to produce a ready-to-use permanent store of vessel information for maritime intelligence. This is particularly useful to analyze past behaviors of vessels, and hence evaluate an index risk to be associated to each vessel.

REFERENCES

[1] Eurostat Report DS-022469, “International trade in goods by mode of transport”, data extracted in May-June 2017.

[2] R. O. Lane, D. A. Nevell, S. D. Hayward and T. W. Beaney, "Maritime anomaly detection and threat assessment," 2010 13th International Conference on Information Fusion, Edinburgh, 2010, pp. 1-8.

[3] Filipiak, Dominik & Stróżyna, Milena & Węcel, Krzysztof & Abramowicz, Witold. (2018). Anomaly Detection in the Maritime Domain: Comparison of Traditional and Big Data Approach. 10.14339/STO-MP-IST-160.

[4] Paul Zikopoulos, Chris Eaton. 2011. Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data (1st ed.). McGraw-Hill Osborne Media.

[5] James Murty. “Programming Amazon Web Services: S3, EC2, SQS, FPS, and SimpleDB”. O’Really, 2008.

[6] James N. Hughes, Andrew Annex, Christopher N. Eichelberger, Anthony Fox, Andrew Hulbert, Michael Ronquest, "GeoMesa: a distributed architecture for spatio-temporal fusion," Proc. SPIE 9473, Geospatial Informatics, Fusion, and Motion Video Analytics V, 94730F (21 May 2015).

[7] Mehul Nalin Vora, "Hadoop-HBase for large-scale data," Proceedings of 2011 International Conference on Computer Science and Network Technology, Harbin, 2011, pp. 601-605.

[8] I. Lytra, M. Vidal, F. Orlandi and J. Attard, "A big data architecture for managing oceans of data and maritime applications," 2017 International Conference on Engineering, Technology and Innovation (ICE/ITMC), Funchal, 2017, pp. 1216-1226.

[9] L. Cazzanti, A. Davoli and L. M. Millefiori, "Automated port traffic statistics: From raw data to visualisation," 2016 IEEE International Conference on Big Data (Big Data), Washington, DC, 2016, pp. 1569-1573.