Big Data Analysis of AIS Records to Provide Knowledge for Offshore Logistic Simulation

Maricruz A. F. Cepeda, COPPE, UFRJ, Rio de Janeiro/Brazil, maricruzcepeda@oceanica.ufrj.br
Rodrigo Uchoa Simões, PENO, UFRJ, Rio de Janeiro/Brazil, rodrigo.usimoes@poli.ufrj.br
João Vitor Marques de Oliveira Moita, PENO, UFRJ, Rio de Janeiro/Brazil, joaov@poli.ufrj.br
Luiz Felipe Assis, COPPE, UFRJ, Rio de Janeiro/Brazil, felipe@peno.coppe.ufrj.br
Luiz Antônio Vaz Pinto, COPPE, UFRJ, Rio de Janeiro/Brazil, vaz@oceanica.ufrj.br
Jean-David Caprace, COPPE, UFRJ, Rio de Janeiro/Brazil, jdcaprace@oceanica.ufrj.br

Abstract

Today, Big Data is getting popular in shipping where large amounts of information is collected to better understand and improve logistics, emissions, energy consumption and maintenance. In shipping, the Automatic Identification System (AIS) records millions of information of ships operations. However, to get the most of these big chunks of information specific technologies should be applied to process these data within an acceptable time. This paper presents a model to extract patterns from AIS records in the field of supply chain of offshore platforms. Here a solution using distributed processing framework based on Hadoop Hive queries (map/reduce) and Hadoop Distributed File Systems (HDFS) is developed. First, a short benchmark study is present to compare performance of Big Data technology in front of former technology. Second, results of the pattern extraction regarding navigational behaviour of Platform Supply Vessels are presented. Then, the new knowledge is introduce in a stochastic simulation to mimic the supply chain management of offshore platforms. The results shown that the proposed methodology is efficient to reproduce offshore logistic activities taking into account uncertainties related to operational matters, as well as weather uncertainties that affect the system. Moreover, big data technologies are greatly reducing time to extract pattern from considerable amount of data.

1. Introduction

1.1. Contextualization

Offshore Oil and Gas (O&G) industry is one of the most important industries in the world with a direct impact on the worldwide economies. According to annual world energy statistics, it is stated that in 2014 approximately 55% of total energy consumed in the world has been produced from oil and natural gas, *IEA (2016)*. World oil demand grew more strongly in 2015 (+1.8 million barrels per day (mb/d)), *UK (2016)*. The profitability of O&G development activity depends on both the prices realized by producers and the cost and productivity of present and newly developed wells. Prices, costs, and field's productivity have all experienced significant changes over the past decade, *EIA (2016)*. The collapse of oil prices in late 2014 forces the offshore oil key players to reduce their logistics costs to recover competitiveness.

Most of Brazil oil reserves are nestled in offshore fields, a fact that has led the O&G activities to achieve increasing depths. The logistic of Brazilian pre-salt fields are challenging due to the considerable distance to coastline (~300 km).

Every day, 2.5 quintillion bytes of data are created. Datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyse. This is known as Big Data, *MGI (2012)*. Big data is present in key sectors as health care, public sector administration, global personal location data, retail, manufacturing, social personal and professional data, *Zicari (2014)*.

Big data have revolutionized the industry over the past several years. Companies across the various travel and transportation industry segments as airlines, airports, railways, freight logistics, hospitality and others have been handling large amounts of data for years. In addition, today's advanced analytics technologies and techniques enable organizations to extract insights from data with previously unachievable levels of sophistication, speed and accuracy, *IBM (2014).*

Today, Big Data is getting popular in shipping where large amounts of information is collected to better understand and improve logistics, emissions, energy consumption and maintenance.

Using satellite navigation and sensors, trucks, airplanes or ships can be tracked in real-time. In shipping, the Automatic Identification System (AIS), which is used for preventing collisions at sea and by vessel traffic services (VTS), records millions of information of ships operations, between vessels and between vessels and shore facilities. Historical AIS data is a valuable data source used for vessel traffic analyses, port calling information, risk assessment and accident investigation. It may also provide basement for decisions in offshore logistic.

1.2. Gap

Over the past few years, there has been a fundamental shift in data storage, management, and processing. Companies are storing more data from more sources in more formats than ever before. This is not just about being a "data package" but rather building products, features, and intelligence predicated on knowing more about this information, *Sammer (2012).*

Organizations discovers new ways to use data that was previously believed to be of little value, or far too expensive to keep, to better serve their clients. Sourcing and storing data is a part of the problem. Processing that data to produce information is fundamental to the daily operations of any industry, *Sammer (2012).*

However, to get the most of these big chunks of information specific technologies should be applied to process these data within an acceptable time and resources. In order to efficiently extract value from these data, the offshore oil key players need to find new tools and methods specialized for big data processing.

1.3. State of the art

To deal with this huge quantity of data, common solutions may not be any more efficient or too time consuming. Here a solution using distributed processing framework based on Hadoop Hive queries (map/reduce) and Hadoop Distributed File Systems (HDFS) is developed. Hadoop is a useful platform used in the last years where solutions have been developed for the industry with the use of big data.

In shipping industry, the use of big data is present in various studies. *Bons and Wirdum (2016)* describe how CoVadem introduces a big data solution that will add significant value to the inland shipping industry in Europe with a cooperatively sourced big data from over 50 vessels (over 55.000.000 measured values a day). It is used to provide effective key performance indicators (KPI's) to judge actual performance and cater for the necessary metrics to analyse, interpret and decide upon improvement measures. With the right technical and organizational implementation a revolutionary basis is introduced that allows for effective, continuous and holistic improvement, *Bons and Wirdum (2016).*

Rødseth et al.(2016) present an overview of some of these issues and possible solutions about the constraints to the use of big data. New protocol standards may simplify the process of collecting and organizing the data, including in the e-navigation domain are reviewed. This paper references the external ship monitoring as AIS and VTS. It indicates that AIS receivers can provide very valuable data on ship movements due to ships will send AIS data quite frequently, normally minimum each 10 seconds. Data that is transmitted is position, speed, course, true heading and rate of turn. This automated ship reporting is a prioritized solution in the e-navigation strategic implementation plan, *IMO (2014),* so one may expect some developments in this area in the coming years, *Rødseth et al. (2016).*

Ramsden et al. (2016) use a Big Data solution to predicting fouling on an underwater hull because they consider that as a vital part of optimising the efficiency of a maritime vessel. They mash together multiple data streams for relevant vessel attributes, positional data, environmental data and fouling coating performance factor generated a dataset of over 3.5 Billon records, *Ramsden et al. (2016).*

The creation of realistic scenarios with simulation in maritime industry is important for training and for testing of the developed tools. These scenarios are verified and validated with the reality. In the shipping industry the use of simulation is present in various studies. *Korte et al. (2012)* present the project Safe Offshore Operations about offshore training simulations where the main idea consists of the development of an integrated operator assistance and information system based on a self-organising wireless computer and sensor network, and also prolong the time frames for offshore operations. The initial simulate scenarios and their verification shall be presented and discussed by the study, *Korte et al. (2012).*

Shyshou et al. (2010) proposed a simulation model for offshore anchor handling operations related to movement of offshore mobile units. The operations are performed by anchor handling tug supply (AHTS) vessels, which can be hired either on the long-term basis or from the spot market. The stochastic elements are weather conditions and spot-hire rates. The requirements on the weather conditions are similar to the methodology developed in this paper. However, the authors are using met-ocean data to assess theses distributions, *Shyshou et al. (2010).*

The work developed by *Aneichyk (2009)* covers the designing of a simulation model for offshore supply process with the aim of creating a tool to plan the operations and fleet size. Some uncertainty factors affecting the process are taken into account such as weather conditions, varying demand and delays. A discrete-event simulation (DES) model is developed in order to model those uncertainties with a stochastic approach. Results obtained show that simulation may be seen as an important tool to develop new strategies under varying conditions and improve the efficiency of the process dramatically. The authors stated that simulation has a promising future in offshore logistics field and the usage will increase in near future, *Aneichyk (2009).*

The novel approach to use AIS data to feed DES input is original and the problem has not been previously studied in that way.

1.4. Purpose

This paper presents a model to extract patterns from AIS records in the field of supply chain of offshore platforms. Big data technologies are greatly reducing time to extract pattern from considerable amount of data. Then, DES methodology is used to simulate the offshore logistics. The results shown that the proposed methodology is efficient to reproduce offshore logistic activities taking into account uncertainties related to operational matters, as well as weather uncertainties that affect the system.

2. Methodology

Apache Hadoop is a platform that provides pragmatic, cost-effective, scalable infrastructure for building many of the types of applications described earlier. It is an open source software that allows for the distributed processing of large data sets across clusters of computers using a simple programming model. It scales up from single servers to thousands of machines, each offering local computation and storage, *White (2009).*

The Hadoop library is using a filesystem called Hadoop Distributed Filesystem (HDFS). HDFS was built to support high throughput, streaming reads and writes of extremely large files. *Sammer (2012).* HDFS uses a scale-out model based on JBOD (“Just a bunch of disks”) to achieve large-scale storage. It uses replication of data to accomplish availability and high throughput.

Fig.1 presents the differences between the traditional architecture of a development program to process data, and a Hadoop platform. The traditional architecture have a Both Storage Area Networks (SAN) or a Network Attached Storage (NAS). SAN is a local network of multiple devices that operate on disk blocks while NAS is a single storage device that operates on data files. These SAN or NAS are connecting to a database with a bunch of applications connected to it and data is being constantly moved to where processing needs to happen. Such a model does not provide high performance at scale. What you really need is a distributed storage as well as processing platform such as Hadoop, where the functionality is run locally on the data, and the system scales linearly to extreme limits - even to geographically dispersed locations, Fig.1, *Srivas (2017)*.

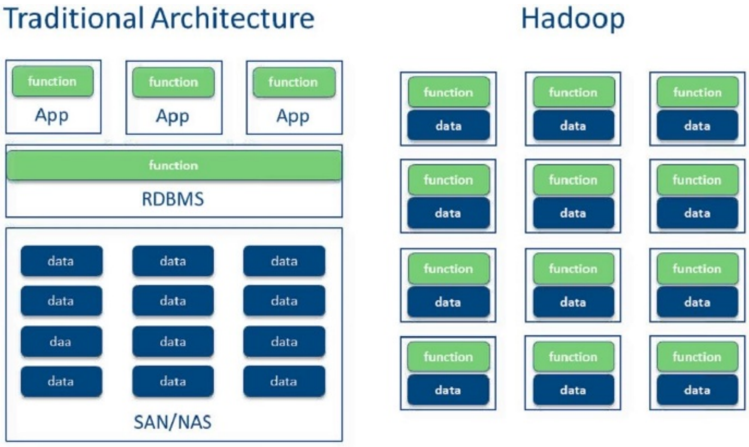


Fig.1: Differences between traditional and distributed architectures (Hadoop), *Srivas (2017)*

MapReduce is a particular programming model for data processing, it is suitable for non-iterative algorithms where nodes require little data exchange to proceed (non-iterative and independent), Fig.2. In MapReduce, developers write jobs that consist primarily of a map function and a reduce function, and the framework handles the gory details of parallelizing the work, scheduling parts of the job on worker machines, monitoring for and recovering from failures, and so forth, *Sammer (2012), Capriolo et al. (2012).*

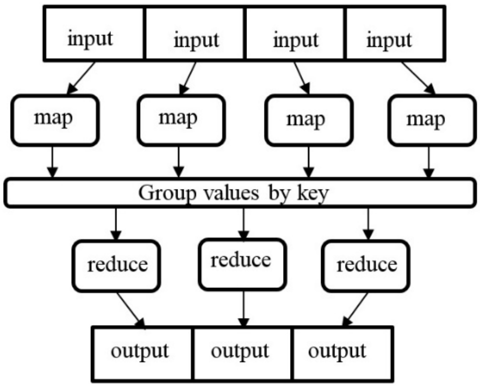


Fig.1: Map-reduce structure

Hadoop structure emerged as a cost-effective way of working with big data. MapReduce breaks up computation tasks into units that can be distributed around a cluster of commodity, server class hardware, thus providing cost-effective, horizontal scalability. Hive provides an SQL dialect, called Hive Query Language (abbreviated HiveQL or just HQL) for querying data stored in a Hadoop cluster similarly to traditional relational databases and the Structured Query Language (SQL), *Capriolo et al. (2012).* SQL knowledge is effective and reasonably intuitive model for organizing and using data. Mapping these familiar data operations to the low-level MapReduce Java API can be daunting, even for experienced Java developers. Hive does this work for you. It translates most queries to MapReduce jobs, thereby exploiting the scalability of Hadoop, while presenting a familiar SQL abstraction, *Capriolo et al. (2012).*

2.1. Benchmark on big data

2.1.1. Database

The benchmark has been performed using AIS messages database designed to store all the types of AIS messages, Fig.3. There is 196 different fields where the most important are MMSI, Channel, Message type, Navigation status, Rate Of Turn, Speed Over Ground, Longitude, Latitude, Course Over Ground, True Heading, Timestamp, Call sign, Name, Cargo type, Vessel dimensions, Estimated Time of Arrival, Maximum static draft, Destination, Data terminal ready, etc.

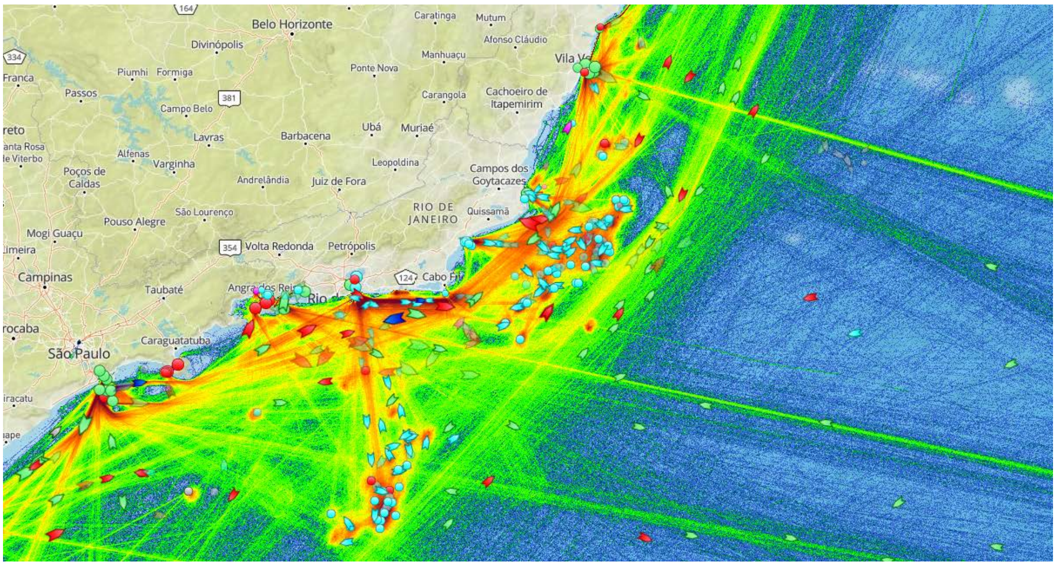


Fig.2: AIS data recorded during 2016 around Campos and Santos basin

The traditional architecture were designed using a Desktop Dell XPS 8700 Intel Core i7 with a HD of 2 TB and 16 GB of RAM. A SQL Microsoft server professional were install on the hardware mentioned in order to perform the SQL query that is the object of the benchmark. The distributed and scalable architecture were designed using 5 different nodes deployed with Cloudera. The Cloudera manager node were connected to 4 Cloudera agents that represent a total of HDFS of 7.2 TB.

Table I: Characteristics of the database included the numbers of records, volume of data, and process time in the traditional and distributed architecture.

ID	**Records**	**Data**	**Tradicional architecture**	**Distributed Architecture**		
	Number	*Volume*	*Seconds*	*#Mapper*	*#Reducer*	*Seconds*
1	1000	317 KB	< 1	1	1	25
2	10000	3.162 MB	< 1	1	1	28
3	100000	31.63 MB	1	1	1	28
4	1000000	312.76 MB	3	2	2	28
5	10000000	3.11 GB	11	12	13	29
6	25000000	7.62 GB	77	29	30	37

To make the comparison a simple SQL query has been developed to count how many AIS messages of each type are presented in the database: “`SELECT MESSAGE_ID, COUNT(MESSAGE_ID) FROM aismessages GROUP BY MESSAGE_ID;`”. This query has been executed on both traditional and distributed architecture considering various size of the database varying from 1000 to 25 million records. The result of the benchmark are presented in Table I and Fig.4. For the considered simple SQL query and small databases inferior to one million records the traditional architecture is more efficient that the distributed architecture. However, for bigger databases the distributed architecture is becoming more efficient. For the specific case of 25 million records, the distributed architecture took half time of the traditional architecture. Considering the amount of data to be treated in the second part of the paper, i.e. 100 million records (~30 GB), it has been decided to use the distributed architecture.

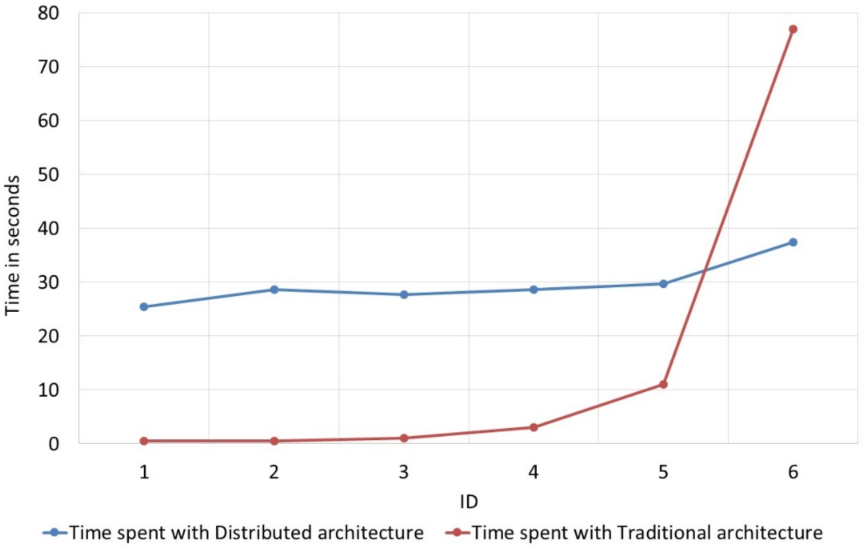


Fig.3: Differences in time spent between a traditional architecture and distributed architecture (Hadoop Hive)

3. Case of study

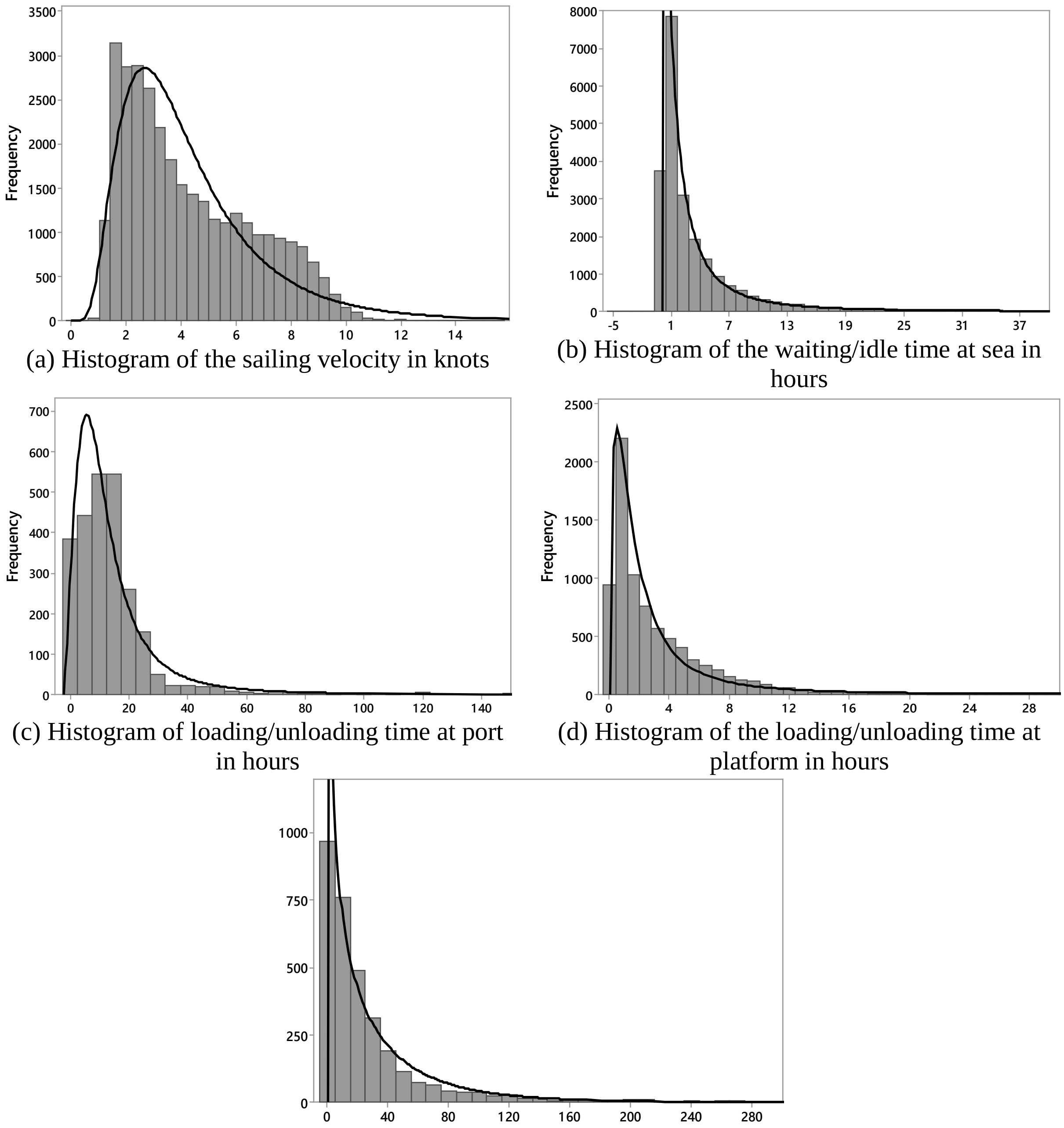
3.1 Extraction of the Statistical Distributions

The case study presented in this section focus on the Campos basin with latitude between 21° 49' 33.2414" S and 22° 40' 47.6969" S and longitude between 39° 42' 26.1104" W and 40° 43' 5.083" W. Several years of AIS data are available in Campos/Santos basin, but in this paper the emphasis is put on a database of 6 months from 1 April 2014 to 1 October 2014. We focused the study on the analysis of 90 platform supply vessels (PSVs) that are performing the supply of the platforms, FPSO's and drilling ships in this region. Finally, these data represents about 100 million of AIS records.

Applying the methodology described in the previous section the histograms presented in Fig.4 have been extracted for the 90 PSVs. Table II gives number of tracks, mean and standard deviations for each behavior of the PSVs while Table III present the probability density functions that best fits the histograms presented in Fig.5.

These data have been generated for any type of PSVs operated during the above-mentioned period. However, in a near future, it is planned to make the difference between various sizes of PSVs as well as to check the effect of weather seasonality on the statistical distributions. Similarly, the histogram related to the loading and unloading time at the logistic port corresponds to various terminals and the loading and unloading time at platform correspond to both production and drilling platforms.

It is interesting to note that the sailing velocity histogram seems to be decomposable into two different components. A first component for the small velocities around 2 knots and a second component for the highest velocities between 6 to 10 knots. This behavior could be explained by the fact that these ships has more than one operational profiles, e.g. PSVs operating close to the platforms are using a small velocity while in transit to port velocity is higher.



(a) Histogram of the sailing velocity in knots

(b) Histogram of the waiting/idle time at sea in hours

(c) Histogram of loading/unloading time at port in hours

(d) Histogram of the loading/unloading time at platform in hours

(e) Histogram of the mooring time at port in hours

Fig.4: Histograms of the 6 months operation of 90 PSVs in Campos and Santos basin

Table II: Extracted data from the AIS database

Behavior of the PSVs vessel	Unit	Number of Tracks	Number of tracks outliers	Mean	Standard Deviation
Sailing velocity	Knots	32 016 (43%)	39	4.34	2.31
Waiting/Idle time at sea	Hours	27 399 (36%)	0	3.75	7.76
Loading/Unloading time at port	Hours	3018 (4%)	5	16.09	33.53
Loading/Unloading time at platform	Hours	9064 (12%)	7	3.24	4.36
Mooring time	Hours	3798 (5%)	0	28.06	49.40

Table III: Best fitting statistical distributions

Behavior of the PSVs vessel	Unit	Type	Parameters
Sailing velocity	Knots	Log Normal	Location = 1.34; Scale = 0.54; Threshold = -0.067
Waiting/Idle time at sea	Hours	Log Normal	Location = 0.44; Scale = 1.33; Threshold = -0,0012
Loading/Unloading time at port	Hours	Log Logistic	Location = 2.51; Scale = 0.46; Threshold = -2,072
Loading/Unloading time at platform	Hours	Log Logistic	Location = 0.54; Scale = 0.72; Threshold = 0,016
Mooring time	Hours	Gamma	Shape = 0.58; Scale = 47.68; Threshold = 0,016

3.2 The Discrete-Event Simulation

To illustrate the concept presented in this paper, a case study based on the Brazilian Campos basin has been developed using discrete-event simulation (DES). The aim of the model is to assess alternative fleet size configurations taking into consideration uncertainty in weather conditions and unexpected delays. The configuration of the layout consist in one logistic port terminal (Macaé) containing 6 berths to load and unload the 23 PSVs considered in the simulation. Fig.6 presents the relative location of the 38 platforms organized in 19 clusters (group of platforms) as well as the port. The DES results are presented for 6 months of the operation of the PSV fleet.

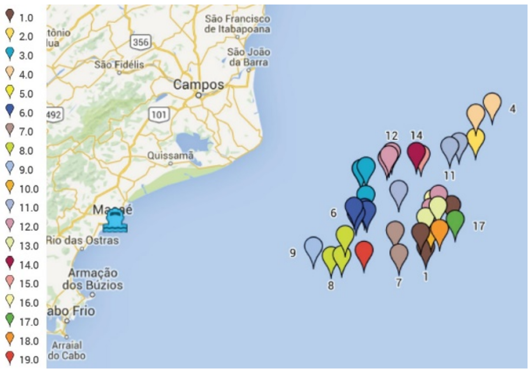


Fig.5: Location of platforms and logistic port terminal (Macaé) considered in the DES. Color of the points represents the 19 clusters of the platforms

In the simulation, platforms are requiring supplies on a periodically base to the logistic port. That period has been defined as a constant in the simulation but differ with the type of the platform. Here the frequency has been chosen around twice a week, i.e., a platform requires a visit of a PSV twice a week. Then, the requests are organized by priority of the load and platform clusters. Depending of the availability of a berth in the terminal, a PSV available in the mooring area is called to be loaded in the port terminal. Finally, the PSV will sail to supply the platform cluster, deliver the load to each platform of the cluster, and then, sail back to the mooring area after the process. The process of PSV allocation and routing is repeated periodically until the end of the simulation.

The dimension, capacity and speed has been considered equal for all PSVs. Moreover, the route of the vessel has been considered straight lines. The statistical distribution of sailing velocity, waiting/idle time at sea, loading/unloading time at port and loading/unloading time at platform has been implemented in each relative process inside the DES while mooring time at port can be considered as the variable to be calibrated.

Fig.7 presents the overall traveled distance by all PSVs along 200 iterations for a period of six months. As the result is highly stochastic, testing the convergence of the output is required. Fig.8

presents the convergence of the overall travelled distance of all PSVs. It is observed that the accumulated mean value tends to converge roughly after 200 iterations, i.e. with a variation of less than 50 km per iteration.

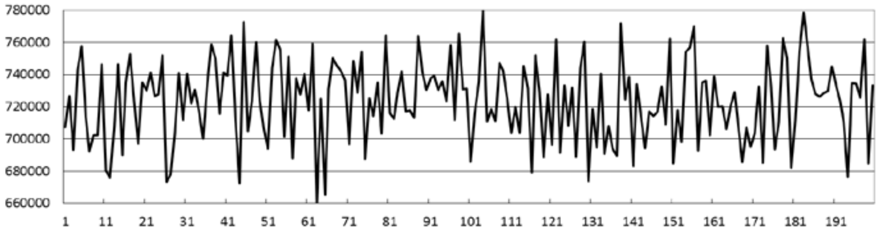


Fig.6: Simulated traveled distance of PSVs during 6 months of operation. Results of first 200 iterations.

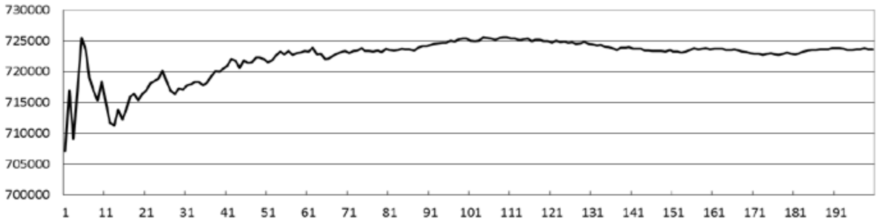
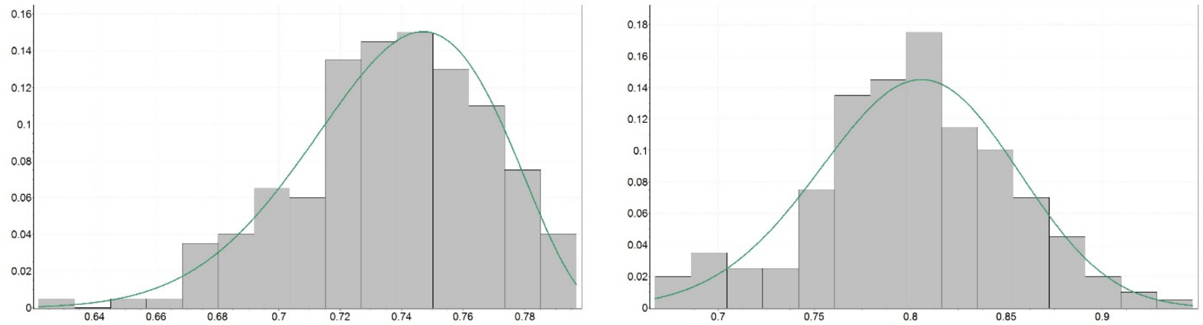


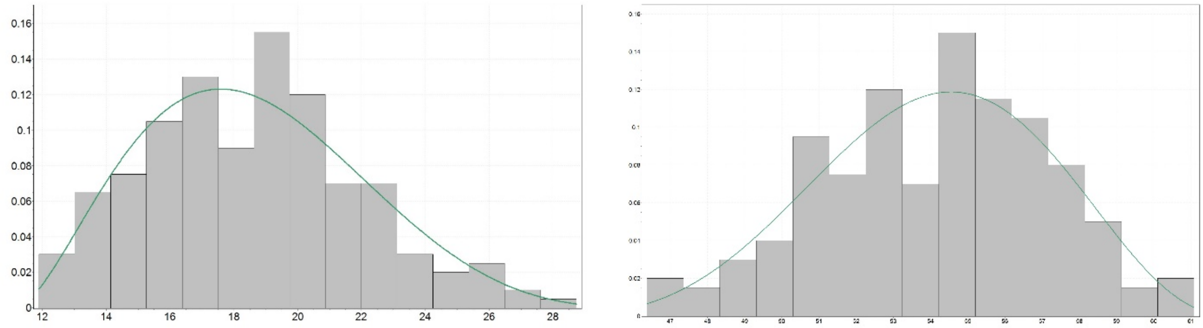
Fig.7: Convergence of the simulation after 200 iterations of the accumulated overall travelled distance measured in kilometers for 6 months of operation.

Any simulation required to be carefully validated. In this paper, the validation data were not presented for confidentiality reasons. However, following section shows typical preliminary results that can be obtained from the DES.



(a) Average utilization of the 23 PSVs in % (b) Average utilization of the 6 berths in %

Figure 8. Probability density function of the utilization of the resources considering 200 iterations and 6 months of operation



(a) Average mooring time per PSV in hours (b) Number of clusters supplied per PSV

Fig.9: Probability density function of typical outputs considering 200 iterations and 6 months of operation

Fig.9 shows respectively the probability density function of the average utilization of the 23 PSVs (mean 75%) and the average utilization of the 6 berths at the logistic terminal (mean 80%). These values indicate that for that configuration the logistic port is saturated. The average mooring time per PSV, Fig. 10 (a), and the number of clusters visited per PSV, Fig.10 (b) are other typical results that can be generated. The mean value of the number of clusters supplied per PSV is around 55. That is cross checking the assumption on two visits of clusters per week for each PSV.

The actual DES model is limited to the study of the influence of the uncertainties due to weather downtimes and sea-going operation delays. However, in a near future, the model would be improved to included additional factor such as deck-load capacity, dry bulk capacity, load delivery delays, etc. Anyway, the presented methodology provides a novel approach to develop realistic simulation starting from the big data fed by AIS data. This represent the basic framework for fleet size decision making, scheduling optimization, cluster optimization, etc. To obtain better results it is recommended to use even bigger AIS database that cover at least 3 years to better mimic the weather uncertainness.

4. Conclusion and recommendations

The use of Big Data software technology such as Apache Hive in Hadoop makes it easy to read, write, and manage large datasets that reside in distributed storage using SQL. Moreover, it may drastically reduce the execution time of the query if enough nodes are used to process the data. This technology has been used here to extract pattern regarding navigational behaviour of Platform Supply Vessels. Then, the new knowledge was introduced in a stochastic simulation to mimic the supply chain management of offshore platforms. The results shown that the proposed methodology is efficient to reproduce offshore logistic activities taking into account uncertainties related to operational matters, as well as weather uncertainties that affect the system.

Acknowledgements

This research was partially supported by Grant 456288/2013-9 of the Brazilian National Research Council (CNPq).

References

ANEICHYK, T. (2009), *Simulation Model for Strategical Fleet Sizing and Operational Planning in Offshore Supply Vessels Operations,* MSc Thesis, Molde University College

BONS, A.; WIRDUM, M. (2016), *Big Data and (Inland) Shipping A Sensible Contribution to a Strong Future*, 15th Int. Conf. Computer and IT Applications in the Maritime, Lecce, pp.420-429

CAPRIOLO, E.; WAMPLER, D.; RUTHERGLEN, J. (2012), *Programming Hive,* O'Reilly.

DARZENTAS, J.; SPYROU, T. (1996), *Ferry traffic in the Aegean Islands: A simulation study,* J. Operational Research 47, pp.203-216

EIA (2016), *Trends in U.S. Oil and Natural Gas Upstream Costs,* U.S. Department of Energy

IBM (2014), *Big data and analytics in travel and transportation*, IBM Big Data and Analytics, pp.1-12

IEA (2016), *Key world energy statistics*, Int. Energy Agency

IMO (2014), *Annex 7: Draft e-Navigation Strategy Implementation Plan*

KOGA, S. (2015), *Major challenges and solutions for utilizing big data in the maritime industry,* World Maritime University, Malmö

KORTE, H.; IHMELS, I.; RICHTER, J.; ZERHUSEN, B.; HAHN, A. (2012), *Offshore training simulations*, 9th IFAC Conf. Manoeuvring and Control of Marine Craft, pp.37-42)

MGI, M.G. (2012), *Big Data: The next frontier for innovation, competition, and productivity*

RAMSDEN, R.; LELLIOT, P.; THOMASON, J.; ROERMUND, D. (2016), *Project Helm: Insights from AIS, Fouling Control and Big Data*, 15th Int. Conf. Computer and IT Applications in the Maritime, Lecce, pp.439-447

RØDSETH, Ø.; PERERA, L.; MO, B. (2016), *Big Data in Shipping - Challenges and Opportunities*, 15th Int. Conf. Computer and IT Applications in the Maritime, Lecce, pp.361-373

SAMMER, E. (2012), *Hadoop Operations,* O'Reilly

SHYSHOU, A.; GRIBKOVSKAIA, I.; BARCELÓ, J. (2010), *A simulation study of the fleet sizing problem arising in offshore anchor handling operations*, European J. Operational Research 203(1), pp.230-240

SRIVAS, M. (2017), *Architecture matters for production success*, MAPR, https://mapr.com/why-hadoop/why-mapr/architecture-matters/

UK (2016), *Oil & Gas UK's Economic Report 2016*

WHITE, T. (2009), *Hadoop: The Definitive Guide*, O'Reilly

ZICARI, R. (2014), *Big Data: Challenges and Opportunities,* Big Data Computing, Western Norway Research Institute, , pp. 103-130