# CargoStats: Automatic Information extraction from freight data

Ricardo Carvalho

ricardo.de.carvalho@ist.utl.pt

*Instituto Superior Técnico*

*October 2014*

*Abstract*— **The freight transport process is complex and involves innumerous entities, such as the exporter and importer companies, the services providers responsible for the transportation of goods, the ship operators, the port entities and the customs officers. Some documents are generated in this process, one of them being declarations which describe, for example, the type of goods transported, as well as their origin and destination. The data contained in these documents reflects the reality of good imports and exports by sea from and to Portugal, which makes them excellent material source for statistical analysis to allow several entities a greater understanding of the process.**

**In this present project a platform was developed, named CargoStats, which makes use of documents sent by the customs to the Portuguese National Institute of Statistics (INE). The platform includes four modules: the ETL module, responsible for the extraction, transformation and loading of the data; a Data Warehouse, in which the data are stored; the cube, constituted by dimensions and measures that will allow further analysis; and an Excel interface for data visualization. Based on the available temporal series in the Data Warehouse, value previsions were created according to the ARIMA models. The solution was evaluated through validation, precision and performance tests.**

*Key Words* — **Data Warehouse, Freight Transport, Forecast, Microsoft Tools**

## I. INTRODUCTION

The freight transport process is complex and involves innumerous entities, making their understanding essential in the process of grounded decision making when it comes to investments and strategic plans. Importation and exportation are processes that involve innumerous administrative documents, among which the declarations that describe, for example, the type of freight transported, as well as its origin and destination [1].

Nearly 1 million freight containers are imported and exported to and from Portugal, and about 20 thousand freight descriptive documents which portray the Portuguese commercial exchanges with the rest of the world are created in the process, given the fact that nearly 99% of these exchanges with non-belonging European Union (EU) countries are made through sea [2].

In Portugal, in the last years, there has been a progressive computerization of systems that support the processes of importation and exportation, making it increasingly easier to standardize the data collection. However, the analysis and disclosure of this data has not followed the evolution of the systems: little information is publicly available and the one that is is outdated. Moreover, the confidentiality of data becomes an obstacle to the development of solutions that can serve stakeholders.

The present project was developed in a business environment, in MAEIL, and aimed to develop a solution intended to allow entities to realize international trade by sea in Portugal, using data generated in the process of freight transport by sea.

Such solution, CargoStats, consists in a Business Intelligence (BI) platform that uses data from the National Statistics Institute (INE). The development of the platform was done in SQL Server 2012, using the components of Integration, Analysis and Excel for ad-hoc analyzes. The platform enables the integration and analysis of maritime cargo based on four dimensions: flow (importation/exportation), time, type of freight and origin/destination of the. The measures used were the number of transactions, mass and statistical value of the freight transported. Components for a one-year forecasting were added, using data mining techniques from the Microsoft Time Series. In this context some comparisons with other prediction models were made.

This document presents a general description of the platform, as well as its use (input and output data, examples of use, limitations) and technical information (configuration and technical description).

## II. GENERAL DESCRIPTION

The platform CargoStats has 4 main modules – see Picture 1: the ETL module, in which the extraction, transformation and loading of the data from the data sources are made; the Data Warehouse module, which clusters all the data; the analytical module, in which the cube that will feed module 4 is built; and the interface that allows the users to see the information generated in the platform.

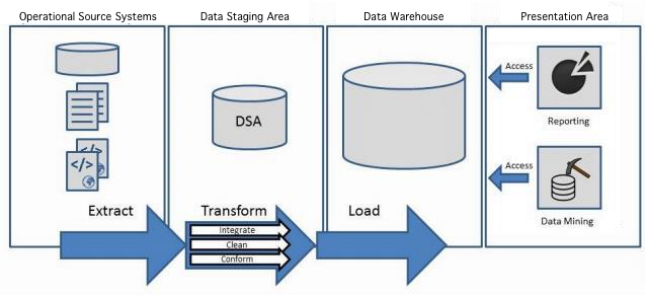


*Fig. 1. General Architecture of the Solution*

The main source of data is microdata from INE for the available years and did not suffer any type of review. However other data shall feed the platform, complementing and completing the transactions. The data sources include [3]:

- Microdata from INE containing transactions from 1992 to 2010;
- Table of classification of freights according to the 2014 Combined Nomenclature – obtained through INE's website;
- Tables of conversion of the classification of the freights between years – these tables allow the perception of the modifications from one year to another – obtained through EUROSTAT's website;
- Freights classification tables according to NST – obtained through INE's website;
- NST classification conversion table to the NC classification – obtained through EUROSTAT's website;
- Countries and regions geographic classification table – obtained through INE's website.

The microdata from INE are obtained through two administrative processes: the INTRASTAT and EXTRASTAT, for commercial exchanges inside and outside the EU, respectively[4].

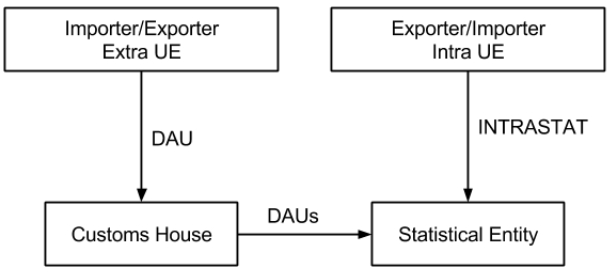


*Fig. 2. Methodology for data acquisition*

The INTRASTAT results in intensive surveys to the importers and exporters. The EXTRASTAT is based on the use of a custom process – the single administrative document (DAU) sent by the importer or exporter to the custom. These documents are then sent weekly to INE [5].

**ETL**

CargoStats allows the importation of data through .csv files, provided by INE, with information about the transaction made: year, month, flow (importation or exportation), commodity classification code, origin/destination country, statistic value and net mass.

The Figure 3 show the mapping between the source of data and the fact table. Some attributes are joined to create the keys. Other are updated: the commodity code is updated to the 2014 classification.

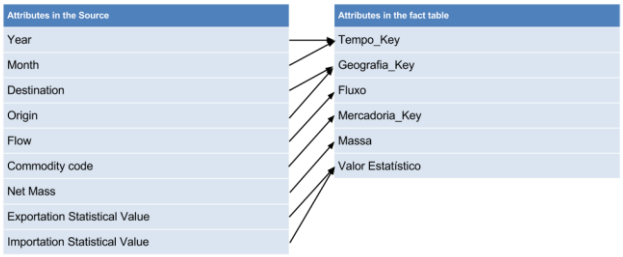


*Fig. 3. Mapping source – fact table*

**Data Warehouse**

The data are stored in a database star schema. It is possible to make queries directly to the database, although no individual data can be provided.

The Geography dimension describes the location from where the cargo was imported or where the exported cargo was issued. This dimension consists of a hierarchy that shows the information can be aggregated to produce different views (by country or region). The Time dimension contains the month, quarter, semester and year of the transaction; provides different views of time as monthly, quarterly, semiannually or annually. The Freight dimension describes the type of freights transported in the transaction. Two classifications are used, so there are two different hierarchies. The flow is a degenerate dimension: it is derived from the fact table and has its own dimension table.

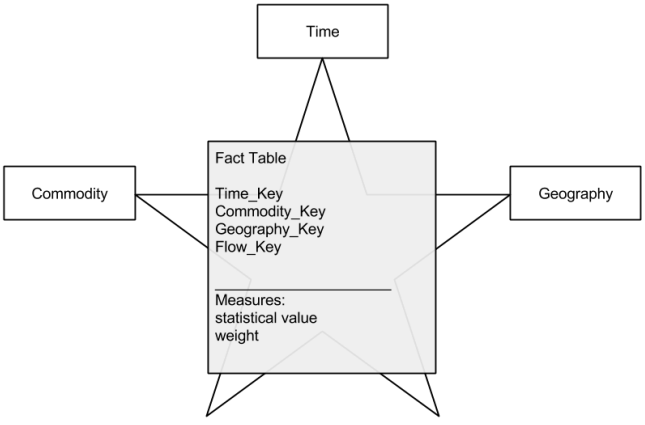


*Fig. 4. Star Model of the DW*

The measures are the net mass corresponding to the mass of the freight itself, devoid of all packaging, in kilograms, and the statistical value that represents the value of the freight at the place and time at which it leaves or arrives at national statistical territory. It is noteworthy that the

statistical value in the importation and exportation has different settings as explained in Table 1 [6].

| *Attribute* | **Description** |
|---|---|
| *Year* | Year of the transaction. |
| *Month* | Month of the transaction. |
| *Destination* | Last country or known statistic territory, at the moment of the dispatch/exportation, to which the freights must be dispatched/exported. |
| *Origin* | Country or statistic territory from which the freights were initially dispatched intended to Portugal, regardless the countries crossed during the transportation. |
| *Flow* | Importation or exportation. |
| *Freight code* | Classification of the freight according to the combined nomenclature. |
| *Net mass* | Actual mass of the freight, devoid of all packaging, in kilograms. |
| *Exportation statistic value* | Value of the freight at the place and time at which it leaves the national statistic territory. |
| *Importation statistic value* | Value of the freight at the place and time at which it arrives in national statistic territory, determined based on the notion of custom value. Equivalent to the CIF value – value of the fright to the exportation, including the expenses to get to its destination (cost of the freight, insurance and shipping). |

*Table. 1. Description of the DW attributes*

**Data analysis**

Based on the data from the Data Warehouse, a cube was created. In the dimensions are available the hierarchies represented in picture 4.

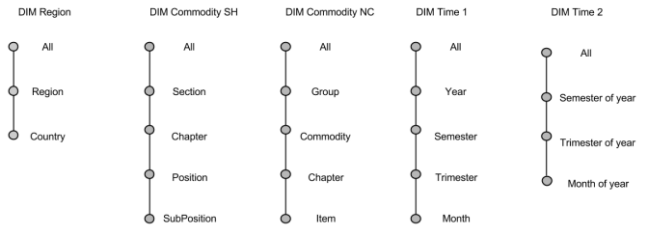


*Fig.. 5. Hierarchies available in the solution.*

Various calculations were created to allow objects that are not defined by the data from the cube to be added. These calculations include:

- Trade balance of freights;
- Mass variations;
- Statistic value variations;
- Percentage of mass variation per geography and freight;
- Percentage of statistic value variation per geography and freight;
- Rate of mass coverage;
- Rate of statistic value coverage;
- Rate of importations per exportations coverage.

**Forecasts**

Data Mining techniques are applied to the historical data in the Data Warehouse. ARIMA models available in the Microsoft Time Series are used.

In statistics and econometrics, and in particular in time series analysis, an autoregressive integrated moving average (ARIMA) model is a generalization of an autoregressive moving average (ARMA) model. These models are fitted to time series data either to better understand the data or to predict future points in the series (forecasting).

The ARIMA methodology makes it possible to determine dependencies in observations taken sequentially in time, and can incorporate random shocks as part of the model. The ARIMA method also supports multiplicative seasonality.

**Interface**

CargoStats allows navigation in the data through an Excel workbook consisting of 5 sheets. The first gives an overview of the data. Three sheets exploring each dimension are then available, and it is possible to select the flow to be explored (importation or exportation):

- Freight, allowing the analysis through the years, as well as the freights more frequently marketed during a certain year and/or in a certain region;
- Region, allowing the analysis of the evolution through the years, as well as the regions and countries with more commercial exchanges with Portugal, per type of freight;
- Time, providing an interface for temporal analysis, including comparison to prior periods.

Additionally, the interface allows the exploration of forecast values for one year – 2011. This feature is available in another sheet where it is possible to have a big perception of the past and forecast values.

**CargoStats does not allow:**

- Access to disaggregated data from the interface - for reasons of confidentiality the user may only have access to aggregate data;
- Selection of different methods of forecast;
- Generation of reports.

**Do you need to use CargoStats?**

Cargostats is an app that allows answer a few questions about:

- Trade orientation in terms of commodities and geography;
- Trade intra-industry;
- Exportation growth;
- Trade intensity in a geographical level;
- Trade complementarity.

The platform was special designed for 3 major entities:

- Public entities interested in trade performance, economic indicators and demand forecast;

- Industry that wants to identify new markets and anticipate changes and tendencies;
- Ports who needs a clearer vision of the trade market to analyze new investments.

**Why is CargoStats different from other apps?**

Cargostats uses data captured from an administrative process that is validated by the National Institute of Statistics. All the data is reliable.

Cargostats is focused in maritime freight transport which allows specific analysis on the field.

It is easily extensible for other means of transportation. The models are easily adaptable to include other attributes, and other data marts can be developed. Other interfaces can be created accessing the data marts.

The main differentiator factor is the variety and actuality of the commodities classifications. There are two major classifications used by the entities and they are available in the Cargostats.

The inclusion of forecasts is another feature important in Cargostats, unprecedented in data from Portugal.

## III. Utilization

CargoStats allows the importation of new data and the use of the interface. The steps that guide these actions in the platform are described next.

**Load of data**

CargoStats receives .csv files which must be placed in the C:/Cargostats/Recebidos folder of the project installation machine. Files are consumed according to a configured time interval (in this case, the default is monthly, since the INE provides files per month). The file format should be .csv, containing each line the information separated by a vertical line, as noted below:

```
Year|Month|Flow|Freight|Country|StatisticValue|Massa
```

An example of a line corresponding to a transaction is the following:

```
1993|01|1|87112099|800|2992|334|334
```

The flow code must be identified with 0 for importation and 1 for exportation. The item code must follow the nomenclature of the Harmonized System. The country code should follow the current naming of INE for the current year. Mass values and statistical value must be numeric.

The data consumed in the C:/Cargostats/Recebidos folder are extracted to the TransaçãoDS table. This table contains all transactions in the format in which they were sent by the INE. After importing all files - end of the extraction process, the process of transformation begins. The data resulting from the transformation process are loaded into a table already in star format - Data Staging Area (DSA). The DSA contains 4 tables: TempoDSA – contains all the information related to the time; MercadoriaDSA – contains information related to the classifications of the freight, including the classification of the Harmonized System and the Combined Nomenclature; GeografiaDSA – table that contains the geographic classification; and FactDSA which contains the facts. The data goes through the following transformations:

- The year and month is transformed in one single attribute and the lookup to the time table is done;
- The article code is transformed in the key format of the Freight table and the lookup to such table is done;
- The flow code is transformed into the keywords Importation or Exportation as appropriate;
- A lookup to the Geography table is done with the code of the country.

Finally, the transaction is added to facts table connected to the respective dimension tables.

The transactions whose lookup result was in the match output are placed in a file Erros.csv in the folder C:/Cargostats/Erros for further analysis and to be replaced in the C:/Cargostats/Recebidos folder. The transactions whose lookup result was match output are loaded in the DSA.

When a stable model is reached in the DSA, the data is loaded into the Data Warehouse tables. The cube is processed whenever the Data Warehouse receives new data - if all goes well, monthly.

**Interface Utilization**

The interface was developed with a focus on the three dimensions of the project: freight, geography and time. The flow dimension can be selected among these, allowing a scan in full or of a particular flow. The main features of the interface are the dimensions, types of representation, navigation means and measures:

- When it comes to the freight, the NST/R classification or the Combined Nomenclature can be selected, varying the level of aggregation available;
- At the geographic level, regions or countries can be selected;
- At the temporal level, month, quarter, semester and year selections are available;
- The flow can either be importation or exportation.

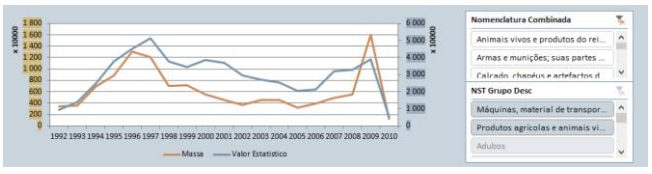


*Fig.6.. Commodity window*

- Temporal graphs - consisting of a horizontal axis representing time, two lines representing statistical values of weight and statistic value, and two vertical axes for each line;
- Bar graphs - in the freight and geography windows, show the distribution of the sum of the values in the highest level of aggregation, allowing an overview of

this distribution, allowing an overview of such distribution;
- Tables - in the freight and geography windows, indicating the three items with higher values for the sum of mass and statistical value in the second maximum aggregation level.

In the Commodity window it is possible to analyze the evolution over time, as well as have a perception of the main freights traded. In the upper area of the window, a graph shows the mass value and statistical value in two axis (left and right, respectively) through the years available (1992 to 2010). It is possible to select other variables such as flow (importation or exportation) and the region through the buttons on the right side of the graph. In the bottom area, a top of freights transacted is displayed, for the sum the statistical value and for the sum of the weight. On the left side, it is possible to analyze the distribution of values for the NST groups that can then be selected in conjunction with the year, to be presented the three with higher statistical value and mass in freights - the total, percentage of the total and variation in relation to the previous year are also presented. The user can also choose to see the table in ascending order of variation, knowing what freights had the highest growth rate in relation to the previous year.

The Geography window is organized in the same way that the freights window. In the upper area of the window it is possible to analyze the evolution of the statistical value and mass over the years (1992-2010), and a filtration by region and/or country is possible. It is also possible to make selection through flow (importation or exportation), as well as the freights. In the lower part, a top of the countries is displayed both for the sum of the statistical value as to the sum of the mass. The right side is possible to analyze the distribution of values across regions that can then be selected in conjunction with the year, to be presented the three with highest statistical value and mass in freights - the total, percentage of the total and variation in relation to the previous year are also presented. The user can also choose to see the table in ascending order of variation, thus to know which were the countries with the highest growth rate in relation to the previous year.

In the Time window it is possible to obtain an annual assessment of international trade transactions by sea. Thus, selecting the year you get a series of data, including monthly evolution graph and a set of values for the mass and the statistical values, corresponding to the total and the variation compared to the previous year, as well as the rate of coverage and variation in relation to the previous year. These values can be filtered by region and group of freights. At the bottom, filters can be used to analyze prior periods, allowing the comparison between months, quarters, and semesters of the years for which data are available for importations, exportations or total transactions.

The user can access the forecasts generated automatically by Microsoft Analysis Services. The table is fed through a connection to the Data Mining structure, which is then modeled with PowerPivot tool.

The forecast window contains a graph through which one can see the evolution of measures over the years; the values of predictions are dashed. Already on the right side, the user can select the flow, freight and region. The graph is updated depending on the choices of the user. The values of the estimates are calculated for the maximum level of disaggregation, and then added when a higher level of aggregation is selected. This window also enables a time selection through the temporal line located at the bottom, allowing the exploration of specific periods.

**Limitations**

- The process to update the commodity classification tables is not that simple – it will be needed to update the classification conversion table for each year;
- It will be necessary to update the commodity classification every year even if it's just few codes;
- Only if the user has skills to manage the powerpivot excel feature it's possible to do ad-hoc queries. The interface available only let the user see, for example, the top 3 importers. To have a bigger list it would be required to know how to do that.

## IV. Technical Information

Cargostats is a platform developed on Microsoft systems. The machine in which the system will be set must have SQL Server 2012 installed, with data tools: Integration and Analysis Services.

Through the Integration Services it is possible to set the folder where the files will be received: accessing the package integraçãoci.dtsx accessing, and selecting the for-each cycle with the name Extração. It is also possible to change the periodicity of data extraction using SQL Server Agent.

Parameters of the forecasts can also be set. To do this simply access the Data Mining structure created in Analysis Services. You can set:

- AUTO_DETECT_PERIODICITY is set to 0 because it's known the periodicity of the data: annual.
- FORECAST_METHOD is set to ARIMA.
- MINIMUM_SERIES_VALUE is set to 0 because none of the values can be negative.

Excel can be installed on any machine that has permissions to access the cube. The excel book has a size of 7GB, so it is necessary to ensure that disk space. The minimum version of Excel is 2013.

## V. Conclusion

The first objective of the project was to identify the current needs of information related to the negotiation process of importation and exportation of freights to and from Portugal. Initially, it was necessary to study the entire universe of shipping and identify key stakeholders in the process. Once in the possession of such knowledge, it was necessary to understand the type of data available to

perform the work. The search and survey the current Portuguese system were essential at this stage. It could be observed that certain questions cannot be assigned in with current systems, although there are data that can be analyzed to answer them.

The study of the development of a Data Warehouse project allowed perceiving the different approaches that can be followed, as well as their advantages and disadvantages. With this project it was possible to show a development process that can be applied to the transport of freights, starting from the case of shipping. In addition to the survey of requirements, where many are transversal to other means of transport, the conceptual design and proposed logic can be extended, serving as a reference.

One of the main challenges of the project was the extraction, transformation and loading. A methodology-solution was presented in the project, to deal with different classifications in terms of both freights as the geographical level. The final work allows the user to select between different classifications used by different entities. It should also be noted that the changes were made such that one has an analysis with uniform codes through the different years. The study of the different classifications and their transformations over the years was essential.

To apply forecasting models to the data from the Data Warehouse allows the end users to have a perception of future trends in commercial trades. Several techniques allowing a comparison of those that best apply to these time series were studied.

## REFERENCES


[1] R. L. Thompson, "U.S. Customs Data: Parsing & Normalization. The first steps in its Long, Transformational journey," 2013. [Online]. Available: http://worldtradedaily.com. [Acedido em 4 12 2013].

[2] Ministério da Economia e do Emprego do Governo de Portugal, "Plano Estratégico dos Transportes: Mobilidade Sustentável - Horizonte 2011-2015," 2011.

[3] DGITA, "Sistema de Tratamento Automático da Declaração Aduaneira," Direcção Geral de Informática e Apoio aos Serviços Tributários e Aduaneiros, Lisboa, 2006.

[4] Instituto Nacional de Estatística, "Sistema Integrado de Metainformação," Instituto Nacional de Estatística, [Online]. Available: http://smi.ine.pt. [Acedido em 24 Julho 2013].

[5] EUROSTAT, "Eurostat's Metadata Server," European Comission, [Online]. Available: http://ec.europa.eu/eurostat/ramon/. [Acedido em 14 Julho 2013].

[6] Instituto Nacional de Estatística, "Documento Metodológico," Departamento de Estatísticas Económicas, Lisboa, 2010.