10-2020

Improving Organizational Decision Making Using a SAF-T based Business Intelligence System

Bruno Oliveira

Mariana Carvalho

Rosa Silveira

Telmo Matos

Improving Organizational Decision Making Using a SAF-T based Business Intelligence System

Bruno Oliveira, Mariana Carvalho, Rosa Silveira and Telmo Matos
CIICESI, Escola Superior de Tecnologia e Gestão, Politécnico do Porto, Portugal,
{bmo, 8180717, mrc, tsm}@estg.ipp.pt

Abstract

Today, companies need to quickly adapt to business changes and react to customers' tendencies and market demands in an unpredictable environment. In this field, the analytical systems represent an important asset that each company should have and use. Data Warehousing Systems (DWS) support companies' analytical needs, however, the development and integration of the data systems is a critical part. Due to specificities of the involved data, each DWS is unique, which compromises the use of reusable components or even the use of pre-built solutions. In this paper, we propose a standard skeleton for a DWS based on Portuguese Audit Tax documents (SAF-T (PT)). These documents represent a standardized procedure for every Portuguese company, providing the necessary data about billing, accounting, and taxation. Thus, they can provide the foundations to use them as a standard data representation to create a DWS that can be posteriorly explored by analytical techniques to generate useful insights.

Keywords: Data Warehousing; Data Mining, E(Extract) T (Transform) L(Load), SAFT-T (PT), Data Warehouse Skeleton

1. INTRODUCTION

Today, many companies have a solid understanding of the importance of business digitization as a way to quickly understand their position in an increasingly competitive market. The need to react to customers' tendencies and market demands, forced organizations to quickly adapt to business changes. However, for small or immature organizations, these systems are hard to implement and manage, which can represent a significant investment without (fast) return. One of the biggest challenges for the implementation resides on the data itself since not all data can be at a maturity level to be used for informed decisions, which can compromise or derail the implementation of the analytical system.

Data Warehousing Systems (DWS) support companies' analytical needs, however, its development can take weeks or months, which has an impact on the financial and strategical levels. The Extract-Transform-Load (ETL) is the most critical component for any DWS/BI system (Ralph Kimball & Caserta, 2004). This system is responsible to extract, transform, conciliate, and load data from data sources (internal and external) to data structures especially configured to support decision-making requirements (Ralph Kimball & Caserta, 2004). Due to the complexity of managing data, these

processes are responsible to consume the main resources to implement Business Intelligence (BI) systems, representing a critical component that compromises system' adequacy: if they fail to provide data quality, the system's trust is compromised (Guo, Xu, Shi, Xu, & Tao, 2019; Ralph Kimball & Caserta, 2004).

The DWS complexity resides on the specificities of the data involved. Even if we have the same business requirements, people have different ways of thinking, which is related to different ways to store, process, and analyse data. Even when considering a standard solution for the implementation of a DWS (which happens in some business areas), data models already built need to be adapted accordingly to the requisites of those business users and their decision-making requirements.

To attenuate the problems related to DWS implementation for small organizations, in this paper we propose a DWS model based on the data used by the Portuguese Standard Audit File for Tax purposes (SAF-T (PT)). These files represent a standardized procedure for every Portuguese company, providing the necessary data about billing, accounting, and taxation that can be used to control companies' activities by Portuguese tax inspection services. Even simple, these files provide the foundations to use a standard data representation to create a skeleton of a Data Warehousing systems that can be posteriorly enriched to support more business requirements. Also, in this paper, we apply Data Mining (Han, Kamber, & Pei, 2012) to extract knowledge of the dataset to demonstrate the potential of the proposed DW.

The remainder of this paper is organized as follows: the next section is dedicated to describing the SAF-T (PT) documents; Section 3 describes the related work considering the SAF-T (PT) documents and the typical analytic procedures applied to the entities represented in SAF-T (PT) documents; Section 4 describes the proposed DWS based on SAF-T (PT) data; Section 5, presents the architecture to support the proposed DWS; Section 6, describes the potential use of exploratory techniques for knowledge extraction and finally, in section 7, conclusions are addressed and future work directions are suggested.

2. THE PORTUGUESE STANDARD AUDIT FILE FOR TAX PURPOSES

The Standard Audit File for Tax Purposes (SAF-T (PT)) is an XML (Extensible Markup Language) vocabulary created to standardize procedures and potentiate the use of data for business control and inspection. These documents collect all companies' relevant data related to billing activities, allowing the interoperability of data of a set of accounting records, in a common and readable format, independent of the billing software used in each company.

According to the Portuguese Institute of Statistics[1], in 2018, 99% of the Portuguese companies were small or medium-sized enterprises (SMEs) and 96% of these were micro-companies, that is, companies that employ less than ten individuals and bill less than two million euros per year. On the other hand, since 2017, the Portuguese Tax Authority[2] (PTA) requires all companies that have organized accounting to monthly submit the Standard Audit File for Tax Purposes (SAF-T (PT)) for validation. The SAF-T (PT) file supports accounting and/or billing applications, considering specific schema rules applied to each one. The billing SAF-T (PT) document is generated in a regular basis (every month) to communicate the billing operations to the PTA, while the accounting SAF-T (PT) file is a more complete file that can be exported by accounting software and must be sent whenever required by the PTA.

SAF-T (PT) application	Label	# Variables	Type
Billing and accounting	Header	16	text, nominal, integer
Accounting	GeneralLedgerAccounts	11	text, nominal, integer, real
Billing and accounting	Customer	14	text, nominal, integer, binary
Billing	Supplier	14	text, nominal, integer, binary
Billing	Product	4	Text, nominal
Billing and accounting	TaxTable	7	text, nominal, date, real
Accounting	GeneralLedgerEntries	28	text, nominal, date, hour, integer, real
Billing	SalesInvoices	62	text, nominal, date, hour, integer, real
Billing	MovementOfGoods	52	text, nominal, date, hour, integer, real
Billing	WorkingDocuments	46	text, nominal, date, hour, integer, real
Billing and accounting	Payments	42	text, nominal, date, hour, integer, real

Table 1: SAF-T (PT) summary.

The Portuguese law decree 302/2016 (Pública, 2016) defined the requirements and a set of taxonomies to be used. Taxonomies are correspondence tables that facilitate the characterization of taxpayers according to the accounting standards used by different taxable entities. The Table 1 summarizes SAF-T (PT) mandatory fields and their data types, framing each document section considering its application.

[1] https://www.ine.pt, accessed Dec 5, 2019.

[2] https://www.portaldasfinancas.gov.pt, accessed Dec 5, 2019.

In this work, we selected the billing document since it is a more common document among companies and provides enough information to support simple, yet critical business perspectives and metrics. Considering the business domain and specific business requirements/activities, some sections are not mandatory. For example, the "payments" and "movement of goods" may not be used in some situations.

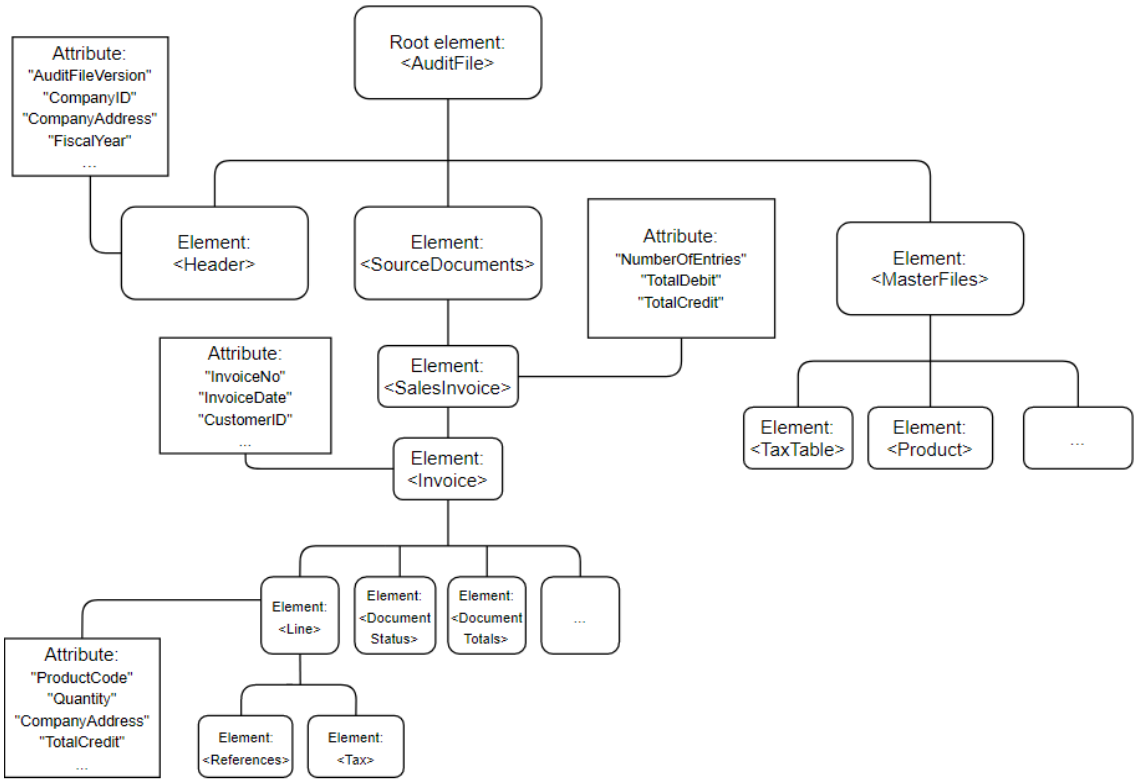

Figure 1 - XML SAF-T (PT) file structure excerpt.

Figure 1 presents a general overview of the main SAF-T (PT) XML elements. The *Header* element describes a general description of the SAF-T (PT) document and the organization that generated it, including the audit file version, the company id, company address, name, fiscal year and software information related to the file generation. The *MasterFiles* element contains the existing information related to the elements that were subject to movement in the tax period to which the document refers. This includes customers, products, services and tax information. Finally, *SourceDocuments* element describes the commercial documents registered in a specific period. This element is hierarchically described by the *SalesInvoice* element that represents the business documents, which include the invoices (*Invoice* element) and the movement of goods (*MovementOfGoods* element). Each one of these elements contains descendent elements describing each type of commercial document. For example, the *Invoices* element contains information about the invoice itself, including the customer, shipping details, or the involved products.

The data represented by billing SAF-T (PT) document is not only valuable for PTA purposes, but also for the companies that generate that data. Despite companies having the tools to export SAF-T

(PT) data, in many cases, they do not have the necessary mechanisms to interpret and extract the meaning of this data in a flexible way. The BI systems allow for that, providing data exploration paths based on data perspectives and metrics identified in the decision support data model. That way, users can explore data beyond the initial pre-defined exploration paths. This provides flexibility and faster insights, giving to the business users more independence for exploring business data.

Based on the SAF-T (PT) files, we propose a BI model that can be applied to every company, providing the foundation to explore core data and at the same time, provide the necessary bridges to extend the initial model with the company´s specific data.

3. RELATED WORK

Using SAF-T (PT) files for BI is not a very common approach. Few studies are found in the literature that explores the relationship between these concepts, or that use these files for decision-making. Rolo et al. (Rolo, Fonte, & Lopes, 2015) proposed a cloud architecture that enables the storage of SAF-T (PT) files in a Data Warehouse (DW). Then, thought a web application, they allow the communication between the government's portals within web services, allowing the automation of tax data submission. Lopes (Lopes, 2016) also proposed an information system, encompassing a reporting tool for the detection of fraud in sales and stocks. This system also allows the control of products inventories for local businesses. Based on SAF-T (PT) files, a prototype and a case study are presented, generated by the cash register that is installed in shops. More recently, Vicente (Vicente, 2017) in its thesis, proposed a reporting-based model on SAF-T (PT) files. They obtain a financial evaluation centred on these files and use the Colibri software dashboard prototype to show the results achieved. A simple decision support system is proposed, mainly using a visual insight into the results obtained with the analysis of SAF-T (PT) documents.

Nowadays, several other software solutions exist (Primavera, SAGE, etc) that generate, autonomously and in a uniform fashion, files to PTA. These solutions gathered business data and monthly send them to these entities. The lack of history preservation regarding the data and sometimes the absence of an analysis of these solutions led to the conception and implementation of a new, more comprehensive solution, thus filling the need of many companies for a new view on their data, from a statistical aspect or if we want to go further, through data mining techniques.

In this work, we simulate a real scenario in which, with the given set of SAF-T (PT) sample files, we applied some Data Mining (DM) techniques to obtain some knowledge about the data itself. DM can be used as a process that allows extracting knowledge and is widely used in decision support, forecasting, description and prediction in many application fields. As we are dealing with SAF-T (PT) data, we can perform a customer segmentation analysis. Some techniques such as Recency, Frequency and Monetary (RFM) analysis are used to assess customer segmentation. Introduced by

RFM is implemented to quantify customer equity and to focus on relationship management efforts. For more detail about customer segmentation and RFM analysis please, refer to (Chen, Kuo, Wu, & Tang, 2009; Dursun & Caber, 2016; Maryani & Riana, 2017). Some other DM techniques can be used to explore data, giving the decision-making tools to drive the organizations forward.

4. A BUSINESS INTELLIGENCE SYSTEM

BI involves complex technologies and strategies that allow end-users to analyse both current and historical data, create reports or performing data-mining tasks. In this context, DWS are being used for a long time ago to store and manage data. Its premise is relatively simple: they represent the organization's single source of truth. This is possible due to a complex work of extracting data for several data sources (typically enterprise sources) and aligning its structure and requirements to the analytical requirements defined by the DW: a single repository specially built to support analytical requirements, providing crucial insights to decision-making processes. Additionally, the DW captures operational data at several points of time, allowing for historic preservation. Preserving historical data is one of the most important aspects of a DW and by consequence the BI system (R Kimball & Ross, 2002). All relevant changes made to source data are preserved, providing an ability to look back in time and provide framed insights to business users.

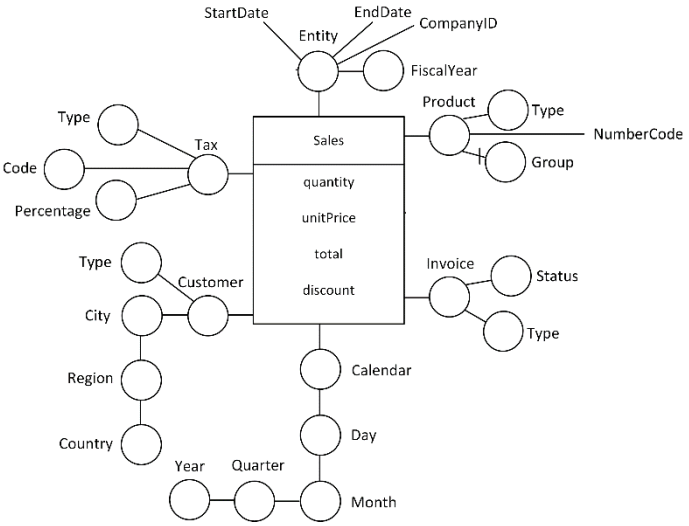

Figure 2 - DM Star Schema used for SAF-T (PT) case study based on billing data.

Even considering a standard data representation for SAF-T (PT), several problems can occur on data instances used to create each document. While structural constraints can be easily validated (using the correspondent XML Schema), the integrity of these data can be difficult to guarantee due to data entry errors, data changes framed on a specific time window, or software bugs that can be hard to detect.

Additionally, history maintenance cannot be guaranteed using only SAF-T (PT) files because each file stores self-contained data about customers, products, tax values, and invoices for a specific

period (a sample of a SAF-T (PT) file can be seen in Figure 1). When analytical procedures are applied, data lineage and evolution are compromised because we have a small picture of the organization at a specific point in time. With a DW system, data that change over time can be captured, providing properly framed data states that can allow for deep data analysis.

To demonstrate the potential DW for SAF-T (PT) related data, we present and explored a case study involving simulated data (essentially due to data privacy reasons) typically found in SAF-T (PT) documents. Thus, step by step, we show how data from billing SAF-T (PT) documents can be migrated to a relational DW schema and how the exploratory techniques can be applied and explored to support typical departmental business activities.

Figure 2 presents a DW conceptual schema using the Dimension Fact Model Notation (Golfarelli, Maio, & Rizzi, 1998). The star schema captures sales events using a fact table (in the middle) and a set of dimension tables proving context and different perspectives for sales data analysis. The Sales fact table represents a sale line event with four measures:

- *Quantity* of a product sold;
- The price of a certain product sold: *UnitPrice*;
- Sale line *Total* obtained by the formula: *Quantity* * *UnitPrice*;
- *The discount* value applied to the sale line.

The star schema also presents six dimensions:

- The temporal data dimension: *Calendar*;
- The characterization of the invoice receiver is supported by the *Customer* dimension, represented by the identification of customer (named or unnamed) type and an address hierarchy: *City* -> *Region* -> *Country* -> ALL. The ETL (Extract-Transform-Load) procedures are responsible to ensure data completeness for region attribute (since it is optional) as well to guaranty data consistency among the several hierarchy levels;
- The *Invoice* data, characterized by the invoice status (states such as a billed document or cancelled document are maintained) and type (holding the information about the several Portuguese invoice types);
- The *Product* data is stored in the *Product* dimension, preserving the *Type* (describing the type of product, considering the Portuguese regulation), *Group* (describing the product family), and the descriptive attribute: *numberCode* (based on EAN[2] codes). The group attribute is optional and does not have a pre-defined set of domain values specified by SAF-T (PT) specification. For that reason, the ETL system should be responsible to

normalize/customize values for this attribute according to the domain specificities (if possible).

- Tax context about each sale, preserving the attributes: *Type* (related to Portuguese legislation), *Code* (represents the tax code considering the Portuguese tax table legislation) and *Percentage* (for specific cases should be defined with 0 value but is generally regulated according to a set of pre-defined rules in Portuguese legislation);
- General information about the sales, including data about the entity that generated the SAF-T (PT) document for the corresponding fact events (*CompanyID*), the sales' fiscal year (*FiscalYear*) and the start (*StartDate*) and end date (*EndDate*) in which the sales are framed.

This is just an example of a possible DW configuration. Additional business processes could be included in this configuration. For example, the "movement of goods" related data could be a good candidate to extend the DW applicability.

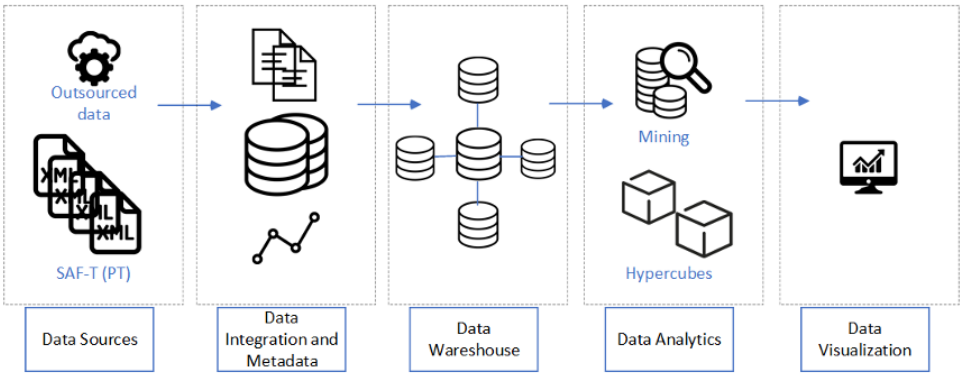

Figure 3: System architecture

5. DATA WAREHOUSING SYSTEM ARCHITECTURE

The DWS architecture typically involves several components to handle data from source systems to the DW repository. The proposed architecture includes a unified service that provides the possibility to submit their SAF-T (PT) documents and based on a set of pre-defined parameters, provides the configuration to enable decision-making support. After document submission, the migration of SAF-T (PT) documents data involves the ETL system. The system works on a temporal basis. It aims to provide a general picture of billing processes done in a specific period. The more information the system contains, the greater is its ability to represent and analyse sales activities for each company.

5.1. The ETL System

The system was organized into four different scalable functional platforms (Figure 3), namely: 1) information sources composed by SAF-T (PT) files and web services used for specific tasks used to enrich source data with external data (the Region attribute from Customer dimension is achieved

recurring to a specific web service specially built for this system); 2) data integration, which is an intermediate system's area holding specific metadata (for example mapping or dictionary tables for data instances normalization) to support the integration procedures; 3) data warehousing, which receives and maintains system's data marts; each data mart supports specific activities of a particular decision-making area. Is this case we are only considering the Sales DW (Figure 2); 4) data analytics, where are located analytical processing mechanisms and structures, providing the means to explore system's data based on the different perspectives of analyses (dimensions). Online Analytical Processing (OLAP) and DM techniques are just two examples of data exploration approaches that can be used; and finally, 5) data visualization, which integrates a set of dashboards that presents the results of the analysis processes.

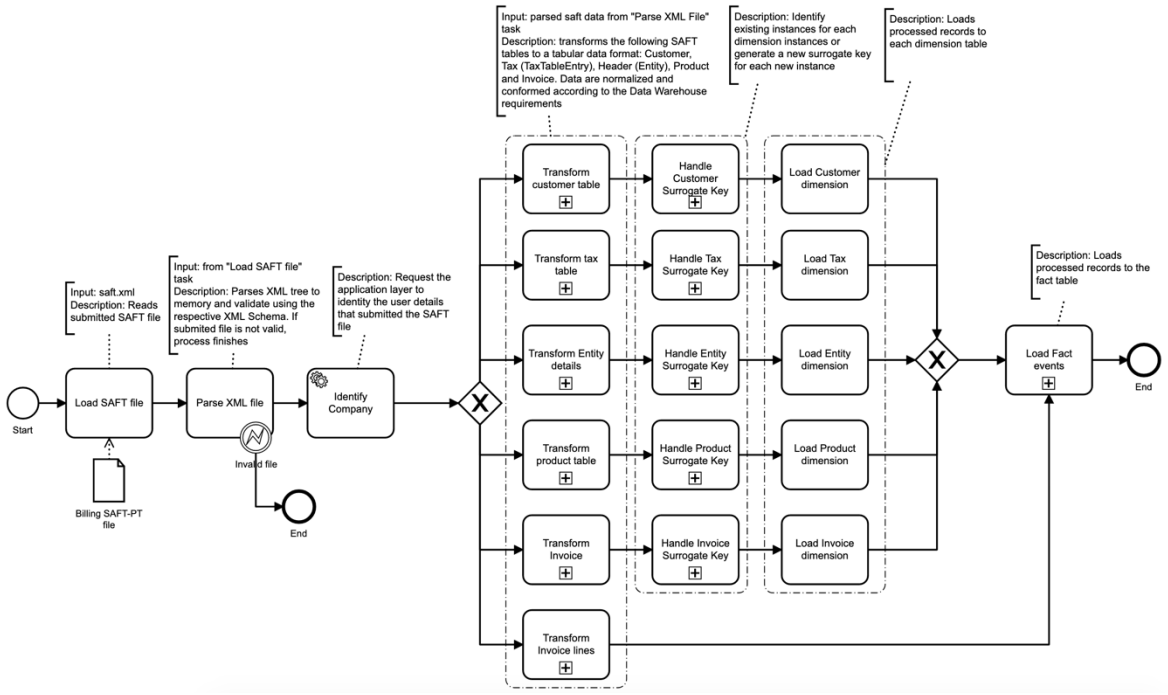

Figure 4: ETL conceptual model to populate the DM Star Schema from Figure 2.

Since each document is processed individually, the amount of data involved in each submission is relatively small. For that reason, each document is processed, and the related data is directly integrated in the DW. The Figure 4 represents an abstract view over ETL configuration using BPMN (Business Process Model and Notation) (El Akkaoui, Mazón, Vaisman, & Zimányi, 2012).

The process starts by reading SAF-T (PT) data (*Load Saft File* task). Then, data are parsed, converting text into an XML DOM object. At this point, data is validated (*Parse XML file* task) according to SAF-T (PT) XML Schema. If the submitted file is not valid, the process will finish at this point since a critical requirement is not valid.

Based on a specific token, an external service is called by the *Identify Company* service task to retrieve the company data that submitted the SAF-T (PT) document. These data are retrieved from the website database used to store the user's data. Next, a set of parallel procedures are executed to

transform data from each SAF-T (PT) used table. Each one of the transformation tasks is identified by a collapsed BPMN subprocess, indicating that each one of them can be decomposed in more granular tasks. In general, these tasks perform a set of typical cleaning and normalization procedures, such as data encoding, trim, normalize lower and upper-case letters, or converting SAF-T (PT) default values for more user-readable values that will be used for posterior analytic reports. For the customer's data, a specific web service is invoked to compose the address hierarchy described before. This specific service holds data about Portuguese addresses and enriches source data form strict hierarchies. The temporal data is not presented in the conceptual model, but a specific service is responsible to generate descriptive temporal data that is not populated on a regular basis. When needed, new records are generated for a long period.

For the SAF-T (PT) tables used to feed each dimension (namely, Customer, Tax, Entity, Product, and Invoice), a Surrogate Key (SK) typical procedure is applied. This new attribute is used as a DW identifier for each record, maintaining the independence between DW and operational data stored in SAF-T (PT) documents. For that reason, before loading each record to a specific dimension, the ETL process verifies if the product already exists. If the product does not exist in the dimension table, then a new SK is generated, preserving the relationship between this new key and the natural key from the SAF-T (PT) document using the metadata layer.

In some scenarios, additional procedures are performed to check data consistency. For example, for products, the *ProductType* attribute is verified to guarantee its value consistency between the stored records and the new ones. When some incoherency is detected, all records (alongside their facts) are stored using quarantine tables that store records that should be posteriorly inspected. These quarantine tables are part of the ETL intermediate layer used to store and manage ETL metadata. Although not covered for the system's current state, Slowly Changing Dimensions can also be used and configured for specific dimensions. In this case, the customer dimension is a typical scenario in which may be relevant to preserve history related to the customer's address.

When all dimension tables are properly loaded, the fact table is populated (*Load Fact events* task), represented through the use of a BPMN collapsed subprocess. This last subprocess implements a Surrogate Key Pipeline process used to replace natural keys from fact records by the correspondent surrogate keys generated along the process.

5.2. System's Metadata

One of the largest challenges in DWS is working with metadata coming from multiple systems to recognize the same logical objects regardless of the possible different representations. Metadata is the data that describes the data. In DWS, the metadata have a critical role at several levels. It can be used to describe attributes in terms of domain or data types, but it is also very important to give name and meaning to the data. Considering the ETL configuration presented, there are important

considerations related to data cleaning and DW extension that can only be achieved considering specific business rules applied in each company business domain.

In (Rahm & Do, 2000), single and multi-source data-quality problems are identified. The single-source problems are partially avoided (mainly at schema level) by SAF-T (PT) specification. Still, some problems may occur and when multiple SAF-T (PT) documents are integrated for the same or different entities. For example, the same product can have different descriptions between different company' entities or products can have contradictory values between different SAF-T (PT) documents. The contradictory values can be identified and handled recurring to quarantine tables, however, the same product with different codes is difficult to handle without additional metadata to help with product identification. For that reason, the use of data dictionary and data mapping tables in ETL systems are important mechanisms to ensure data quality and conformation in the target DW schema.

Multi-source problems can also occur in this scenario. When some companies have several entities (for example, stores) naming conflicts can occur when multiple SAF-T (PT) documents are integrated. This kind of problem occurs frequently at customer and product data instances.

To allow the use of these procedures by each company, transformation rules metadata should be defined in order to allow automatic normalization and corrections during the ETL pipeline.

Another important consideration is related to the DW model extension. As previously referred, the proposed DW approach based on SAF-T (PT) represents only a skeleton. The typical DW data model is characterized by the use of detailed perspectives (dimensions) to provide useful context to the facts for data analysis. The data available in SAF-T (PT) is simple since only a few numbers of attributes can be included. It is the case of *Product* and *Customer* dimensions since they only represent a simplification over the data strucutre typically used. For that reason, an extension mechanism was idealized in order to provide the integration of additional data from the company's source system.

Both for cleaning procedures and data model extension, a set of mapping rules can be provided by each company, allowing for the definition of additional metadata that supports ETL execution. For now, a simple XML file is used to materialize mapping rules. However, as future work, a Domain Specific Language is planned. The main idea is to express the business knowledge to the concepts and logic used by ETL for data enrichment.

6. EXPLORING THE SAF-T (PT) DATA WAREHOUSE

To exemplify the potentialities of the proposed DW and as we are dealing with SAF-T (PT) files, we can perform a customer segmentation analysis using the RFM method (Chen et al., 2009; Dursun & Caber, 2016; Maryani & Riana, 2017). This type of analysis can be performed using clustering

analysis (a data mining technique) that allows for data descriptive analysis. With this approach, we can group customers into clusters and obtain the description (according to its RFMs' characteristics) of the customers.

6.1. *Case Study configuration*

Due to data protection laws, the experimental results were assessed by populating the DW with random SAF-T (PT) documents. This enrichment process was carried out as follows:

- 100 companies and 100 customers were randomly generated across Portugal's regions (NUTS II).
- The fiscal number was randomly generated considering the proportion based in INE[3] and SABI[4] statistics. The first digit classifies the company or customer in the distribution chain: starting with 1 or 2 indicates final consumers or individual companies, those starting with 5 define collective companies and beginning with 9 unknown customer or provisory companies.
- Three different taxes were created, namely 6%, 13% and 23%.
- 1000 products were randomly generated over five product type (P, S, O, E, and I, according to XSD restrictions. This field is mandatory) and randomly distributed over 50 product categories (although this field is mandatory, the entries are free text, that is, the company decides internally its product categories).
- Each SAF-T (PT) has several invoices and line sales connected with each invoice. Invoices are generated considering a date period and a customer, randomly chosen. Each line sale, linked to an invoice, is generated considering: an amount, randomly generated between 1 and 10.000; a product, randomly chosen from 5 and a tax, randomly chosen from 3.

For these preliminary results, the database has 22190 examples, which corresponds to 4438 invoices with 5 lines each for a year. There are 36 companies and each one has a different number of invoices.

6.2. *RFM Analysis*

To perform this customer segmentation analysis, we opt to apply a clustering technique to the dataset. This technique is based on unsupervised learning, in which the class label is not known (unlabelled data) and aims at creating a descriptive model of the dataset, grouping tuples of the dataset in clusters based on the tuples' characteristics.

[3] https://www.ine.pt, accessed Dec 5, 2019.

[4] https://sabi.bvdinfo.com, accessed March 25, 2019.

First, in this process, we start with the data pre-processing phase with data reduction, data transformation, and data normalization. Next, we choose the k-means algorithm, as the clustering algorithm to be applied to the dataset and finally, we validate the outcome of the clustering analysis.

6.3. Data Pre-Processing

The original database contains 84 variables and 22190 examples (invoices lines). We started by reducing the set of attributes by keeping the relevant ones to the RFM analysis (introduced by (Bult & Wansbeek, 1995)): *SKCompany*, *SKInvoice*, *Customer ID*, *Amount* and *InvoiceDate* and considering that only the records of the 8585th company were selected to illustrate the analysis example, the database was reduced to 5 attributes and 47 tuples (customers).

Then, we need to transform the available set of attributes to create the RFM variables. To do this, we need to resort to data normalization. As this analysis deals with different units and scales, normalization is useful to compare attributes values. The standard normalization (Z-transformation method) adjusts the scales of the attributes and preserves the original distribution. Next, we used the attribute *InvoiceData* to define the R variable, or Recency, by calculating the difference between the data of the last purchase with the current time. The F variable, or Frequency, is the total count of *SKInvoices* of the customer. And finally, the M variable, or Monetary, corresponds to the average purchase amounts. Table 2 shows the data statistics of the RFM variables after the data pre-processing phase.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Monetary	47	2920.20	7518.00	4885.3594	938.79433
Recency	47	4.99	16.46	9.2063	3.18958
Frequency	47	5	30	12.65	7.507
Valid N (listwise)	47				

Table 2: Descriptive Statistics.

6.4. Application of the K-Means Algorithm

To perform a customer segmentation via RFM analysis, we can use the k-means which is a data mining clustering algorithm. This is a commonly used algorithm of unsupervised machine learning and it performs clustering on unlabelled data. Unsupervised learning consists in searching for unknown patterns in a dataset with no pre-existing labels. Clustering consists in grouping a dataset or a set of instances into groups or clusters of objects. Each cluster contains objects that are similar to one another and are dissimilar to the objects in other clusters. K-means is a centroid-based technique, which means that the cluster similarity is measured using the cluster centroid, which is an object representing the mean value of the cluster (Han et al., 2012).

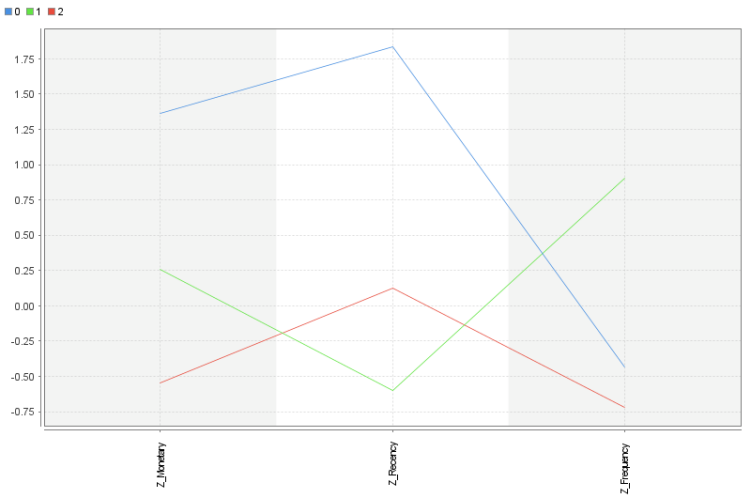

Figure 5 – K-means centroids coordinates from customers.

It is the user's responsibility to choose the *k* number clusters to be created by the algorithm. To choose the best *k*, we can use the Davies Bouldin index (D.L.Davies & D.W.Bouldin, 1979). The goal of this criteria is to minimize the Davies Bouldin index measure as we want to minimize the within-cluster scatter and maximize the between-cluster separation. After some tests, we get the best *k*=3 with Davies Bouldin index equals to 1.052.

Figure 5 shows the outcome of the k-means algorithm. In this figure, we can see the location of each centroid of the three created clusters. There are two centroids with the opposite behaviours. While the type of customers represented by cluster 1 (in green) denotes a good spender (medium monetary values), the last purchase is recent (low recency values) and frequently makes purchases (high-frequency values). The type of customer represented by cluster 2 (in red) denotes a light spender (low monetary values), purchases are not frequent (low-frequency values). Finally, the type of customers represented by cluster 0 (in blue) denotes a heavy spender (very high monetary values) and purchases are not frequent (low-frequency values). This centroids' analysis produces insights that could be used in marketing strategies and allows us to address specific products, campaigns, and optimize marketing costs. For instance, the customer represented by cluster 1 could be classified as a "loyal customer", the one represented by cluster 2 "aversion to shopping" and finally the customer represented by cluster 0 as "outlier behaviour".

According to (Marôco, 2018), the outcome of the clustering analysis must be, *à posteriori*, validated using supervised learning methods. Supervised learning consists in searching for unknown patterns in a dataset with pre-existing labels. So, it uses labelled data to create the model and aims at creating a predictive model. Classification is one of the existing supervised learning techniques. The use of these methods allows us to calculate probabilities of accuracy or error associated with the assignment of objects to a specific cluster made by the k-means algorithm.

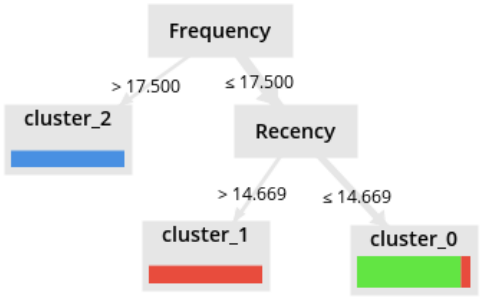

Figure 6 - Outcome of the Decision tree algorithm.

To validate the outcome obtained in the previous section, the 10-fold cross-validation method was used with a classification algorithm, the decision tree algorithm. The decision tree is a heuristic method that uses recursive partitioning. It is a supervised algorithm, and it is performed on labelled data. The decision tree algorithm generates a tree structure, where each internal node denotes a test on an attribute, each branch corresponds to an outcome of the test and each leaf holds a class label (Han et al., 2012). In this case, as we need a class label to perform the decision tree algorithm, the output of the clustering analysis (attribute called "cluster") was used as class label. After performing the validation, we acquire the following tree structure (Figure 6) and the confusion matrix (Table 3).

	True cluster_2	**True cluster_1**	**True cluster_0**	**Class prediction**
Prediction - cluster_2	12	0	0	100%
Prediction - cluster_1	0	20	2	90.91%
Prediction - cluster_0	0	2	13	86.67%
Class recall	100%	90.91%	86.87%	

Table 3 – Confusion Matrix

Figure 6 shows that there are instances (or customers), that have the same characteristics and belong to different clusters. The decision tree algorithm correctly classified 12 instances in cluster 2, 20 in cluster 1, and 13 in cluster 0, of a total of 49 instances. The remaining 4 instances were incorrectly classified: 2 instances (using the clustering analysis) were in cluster 1 and in the decision tree were classified as cluster 0; and the other 2 were in cluster 0 and in the end, were classified as cluster 1. The accuracy value is 91,5%, which means that 91,5% of the instances in the dataset were correctly classified.

7. CONCLUSIONS AND FUTURE WORK

In this paper, a BI system was proposed based on Portuguese tax files SAF-T (PT). This BI system was accomplished through a DWS that provides a DW skeleton to accommodate SAF-T (PT) data. We discussed how this repository can be used for a "self-service" BI and how it can be potentially extended with metadata that allows for a "deeper" data integration and for model extension based

on each company's operational data that do not belong to SAF-T (PT). A Domain-Specific Language is planned as a way to simplify the definition of mappings and rules to business users, which contributes to the independency of business users from TI personnel for such a job.

The proposed architecture includes a unified service that provides the possibility to submit their SAF-T (PT) documents and, with minimal configuration, enables the decision-making support. A case study is presented to demonstrate the DW potential for SAF-T (PT) related data, through the identification of patterns or relationships in the data and exploring properties between them.

To demonstrate the potential of the proposed DW, a customer segmentation analysis using the RFM-method was carried out. Using a clustering algorithm, it was possible to obtain a descriptive model of the dataset and a defined profile of types of customers, which clearly could be used in marketing strategies. Then, the same clustering model was validated with a supervised learning algorithm, a classification algorithm. After the validation, we can conclude that 87,3% instances (or customers) were correctly classified and allocated to the "right cluster", meaning that with the proposed DW, it is possible to perform a customer segmentation analysis and extract knowledge using data mining. Further directions include other data mining techniques, like association rules algorithm with the intent to analyse, for example, relationships between purchased products.

ACKNOWLEDGMENT

This work has been supported by national funds through FCT – Fundação para a Ciência e Tecnologia through project UIDB/04728/2020.

REFERENCES

Bult, J. R., & Wansbeek, T. J. (1995). Optimal selection for direct mail. Marketing Science, 14(4), 378–395.

Chen, Y. L., Kuo, M. H., Wu, S. Y., & Tang, K. (2009). Discovering recency, frequency, and monetary (RFM) sequential patterns from customers' purchasing data. Electronic Commerce Research and Applications, 8(5), 241–251. https://doi.org/10.1016/j.elerap.2009.03.002

D.L.Davies, & D.W.Bouldin. (1979). A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1(2), 224–227.

Dursun, A., & Caber, M. (2016). Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis. Tourism Management Perspectives, 18, 153–160. https://doi.org/10.1016/j.tmp.2016.03.001

El Akkaoui, Z., Mazón, J.-N. N., Vaisman, A., & Zimányi, E. (2012). BPMN-Based Conceptual Modeling of ETL Processes. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 7448, 1–14. https://doi.org/10.1007/978-3-642-32584-7_1

Golfarelli, M., Maio, D., & Rizzi, S. (1998). The Dimensional Fact Model: A Conceptual Model for Data Warehouses. International Journal of Cooperative Information Systems, 07(02n03), 215–247. https://doi.org/10.1142/S0218843098000118

Guo, T., Xu, C., Shi, B., Xu, C., & Tao, D. (2019). Learning from Bad Data via Generation. In 33rd Conference on Neural Information Processing Systems (NeurIPS 2019) (p. 12).

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers (Third Edit). Morgan Kaufmann Series.

Kimball, R, & Ross, M. (2002). The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling (Second). Wiley.

Kimball, Ralph, & Caserta, J. (2004). The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. John Wiley & Sons, Inc.

Lopes, E. M. C. (2016). Novas Tendências , O Big Data. EST - Instituto Politécnico de Castelo Branco.

Marôco, J. (2018). Análise estatística com utilizaçao do SPSS 25. Edições Sílabo, Lisboa (7th ed.).

Maryani, I., & Riana, D. (2017). Clustering and profiling of customers using RFM for customer relationship management recommendations. In 2017 5th International Conference on Cyber and IT Service Management, CITSM 2017. https://doi.org/10.1109/CITSM.2017.8089258

Pública, M. das F. e da A. (2016). Portaria n. 302/2016. Diário da República, 1a Série(231):4273–4379.

Rahm, E., & Do, H. H. (2000). Data Cleaning: Problems and Current Approaches. IEEE Data Engineering Bulletin, 23, 2000.

Rolo, L., Fonte, A., & Lopes, E. (2015). Sistema de gestão e auditoria fiscal na nuvem Management and Tax Audit System in the Cloud. In 15a Conferência da Associação Portuguesa de Sistemas de Informação (CAPSI 2015) (pp. 326–337).

Vicente, L. (2017). Modelo de performance financeiro e comercial para suporte à decisão baseado na norma SAF-T. ISCTE - Instituto Universitário de Lisboa.