Data Requirements Elicitation in Big Data Warehousing

António A.C. Vieira[1] [0000-0002-1059-8902], Luís Pedro[2], Maribel Y. Santos[3] [0000-0002-3249-6229], João M. Fernandes[4] [0000-0003-1174-1966] and Luís S. Dias[5] [0000-0003-1991-4892]

[1,3,4,5] ALGORITMI Research Centre, University of Minho, Portugal
[1,2,4,5] University of Minho, Campus Gualtar, 4710-057, Braga, Portugal
[3] University of Minho, Campus Azurém, 4800-058, Guimarães, Portugal
[1,5]{antonio.vieira, lsd}@dps.uminho.pt
[2] a70415@alunos.uminho.pt
[3] maribel@dsi.uminho.pt
[4] jmf@di.uminho.pt

Abstract. Due to the complex and dynamic nature of Supply Chains (SCs), companies require solutions that integrate their Big Data sets and allow Big Data Analytics, ensuring that proactive measures are taken, instead of reactive ones. This paper proposes a proof-of-concept of a Big Data Warehouse (BDW) being developed at a company of the automotive industry and contributes to the state-of-the-art with the data requirements elicitation methodology that was applied, due to the lack of existing approaches in literature. The proposed methodology integrates goal-driven, user-driven and data-driven approaches in the data requirements elicitation of a BDW, complementing these different organizational views in the identification of the relevant data for supporting the decision-making process.

Keywords: Big Data, Big Data Warehouse, Analytics, Data Warehousing, Hive, Requirements, Industry 4.0.

1 Introduction

Supply Chains (SCs) are complex and dynamic networks, wherein material and information exchanges occur, driven by demand and supply interactions between players [1]. Their goal is to fulfil customers' orders, at a minimum cost, by efficiently managing involved operations, such as: receipt materials, warehousing costs, production, transportation, among others. Companies try to efficiently manage their SC with different systems like SAP-ERP ("Systeme, Anwendungen und Produkte in der Datenverarbeitung" - German for "Systems, Applications & Products in Data Processing"), MRP (Material Requirements Planning) and other tailored-made solutions. Most of these allow companies to respond to specific problems of a given domain, producing large data sets in various formats, at increasingly higher rates. Yet, companies struggle with the extraction of additional knowledge from these data. Such a context is known as Big Data and is one of the pillars of Industry 4.0 [2].

In alignment with Industry 4.0, and to face the needs of integrating, storing and processing data for decision-support in SCs, a Big Data Warehouse (BDW) system is being developed at a company of the automotive industry. For confidentiality reasons, its name cannot be disclosed. However, and to be possible to understand the complexity

associated to this project, as well as the methodological approach proposed in this paper, some figures can be shared. The company in question is part of an international organization that is present in more than 60 countries. It incorporates around 3 000 associates with an estimated sales volume of 700 million euros. Regarding logistic figures, the company works with 600 suppliers spread around the world, which supply more than 8 000 different raw materials through around 230 000 inbound deliveries, per year. In its turn, this culminates in more than 1 200 different finished goods produced and shipped to around 250 customers around the world through more than 40 000 outbound deliveries per year.

Presently, a prototype of such system is finished and the purpose of this paper is to present the work conducted to identify its data needs. With such system, the company in question can accurately and timely perform data analytics methods and thus have a proactive approach, rather than a reactive one, due to the knowledge that is extracted from the integrated Big Data sets. At this point, the focus is not on the evaluation of the solution's performance, but rather in its feasibility deploying a proof-of-concept based on the integration of state-of-the-art contributions from different research areas.

The remaining of this paper is organized as follows. Section 2 discusses related literature. Section 3 describes the methodology applied to identify the data requirements of the BDW. Section 4 presents data sources and the data model used. Section 5 shows an example dashboard for data analytics tasks, supporting the decision-making process. Section 6 discusses conclusions and future research.

2 Related Work

The application of Big Data Analytics (BDA) in SCs is a recent and active research topic, engaged with the development of proactive mechanisms [3]. Kache et al. [4] consider that, despite the advantages of BDA, it is still in its early steps, regarding its application in SC management. In this regard, Tiwari et al. [5] characterized the available BDA techniques and provided a comprehensive review of BDA applications in SC contexts, between 2010 and 2016. Sanders [6] examined the use of BDA to extract knowledge and improve the performance of SCs of leading international companies. The author proposed a framework, based on lessons learned from experience. In their turn, Zhong et al. [7] analyzed several cases throughout the world and discussed their possible impacts on the decision-making process. The authors also reviewed currently used BDA technologies and identified some challenges, opportunities and future perspectives for BDA. Chen et al. [8] examined how the use of BDA can contribute to the added-value in a SC. Santos et al. [2] presented a BDA architecture, under an Industry 4.0 perspective, in a company of the Bosch Group. The presented architecture collects, stores, processes, analyzes and visualizes data related to quality complains and internal defect costs.

To the best of the authors' knowledge, less attention has been paid to the development of BDW systems, applied to SC of the automotive industry. More specifically, from the examples found in the literature, few are oriented towards SC management and no solution oriented towards SC problems of the automotive industry was found. This idea is also corroborated by Ivanov [9].

3 BDW Requirements Elicitation Methodology

The purpose of this section is to describe the methodology proposed for the data requirements elicitation. After the analysis of the available data sources, the dimensional modelling, a traditional approach for the development of Data Warehouses (DW) [10], was used. This approach is not mandatory for the development of a BDW,

as usually, in Big Data contexts, NoSQL schema-free databases are used, with data models that may change over time. Notwithstanding, the authors opted to start with the dimensional modelling, due to some benefits that were identified. First, it allowed a better understanding of the data, culminating in a clearer view over the metrics, Key Performance Indicators (KPIs), organizational processes, and relevant dimensions of analysis. It also ensured the inclusion of the relevant data sources to the problem and the confidence that no relevant data is excluded. Furthermore, this approach also helped in the design phase, because it helped to define the structure of the used Hive tables. Whilst it is true that the dimensional model could have been ignored, the authors strongly believe that, as mentioned, it was helpful and was also the strategy followed by other authors [11–13].

To develop the multidimensional model, there are 2 main methods: (1) the Inmon's top-down [14] and (2) the Kimball's bottom-up [10]. The first consists in designing the DW as a centric system to respond to queries of all stakeholders. The Kimball method, also known as the multidimensional approach, starts by developing individual data marts and then combining them, using the bus matrix, into a single centric DW. In this work, the Kimball's method was followed. Thus, it was necessary to start the project with the business requirements definition phase [15], [16]. This phase, can, in its turn, be divided in three approaches: supply, user and goal-driven; the latter two can be considered as a single approach, known as demand-driven.

The supply-driven approach, also known as data-driven [15], is a bottom-up iterative approach, in which the user requirements are ignored in the first iteration. It consists in solely analyzing the operational data sources, assuming that it is possible to completely derive the DW conceptual model from the database schemas of the data sources. In the subsequent iterations, the user requirements are considered. This approach is simpler and cheaper (in time and money) than other approaches, because it only depends on the DW designer's skills and on the data complexity [15]. Yet, some barriers to implement this approach can be identified. For instance, if there are many data sources, deriving the DW conceptual model can become a very complex task. Furthermore, the lack of documentation, or domain experts, can also increase the complexity of interpreting the data schemas.

The user-driven approach is a bottom-up approach similar to the requirements definition phase in a software development project [15], [16]. In this approach, several types of sessions or interviews are conducted with specialists of different areas, to elicit their requirements. This is a complex task, because it demands that the views of individuals with different perspectives of the same problem and different sensibilities are combined. Thus, this approach can become time expensive, since business users rarely share a common and clear view over the goals and involved processes. One of the main benefits of this approach is that users are highly involved. According to Golfarelli [15], there are risks in only applying the user-driven approach, since users may leave their positions on a company, or even leave the company itself, increasing the difficulty in analyzing the data.

The goal-driven approach [16] consists in conducting a set of sessions with the top-management, to identify organizational goals. Thereafter, the different visions are merged, thus obtaining a consistent view of the global requirements. In comparison with the user-driven approach, in goal-driven approaches there is no risk of obsolete schemas, since the probability of identifying the most relevant indicators is maximized. Since this approach starts with obtaining a view of the global set of requirements for the central DW, it is often considered a top-down approach. Furthermore, due to its similarities with the user-driven approach, they are usually combined by DW designers and are referred as demand-driven. Combining the three approaches, the methodology depicted in Fig. 1 was adopted.

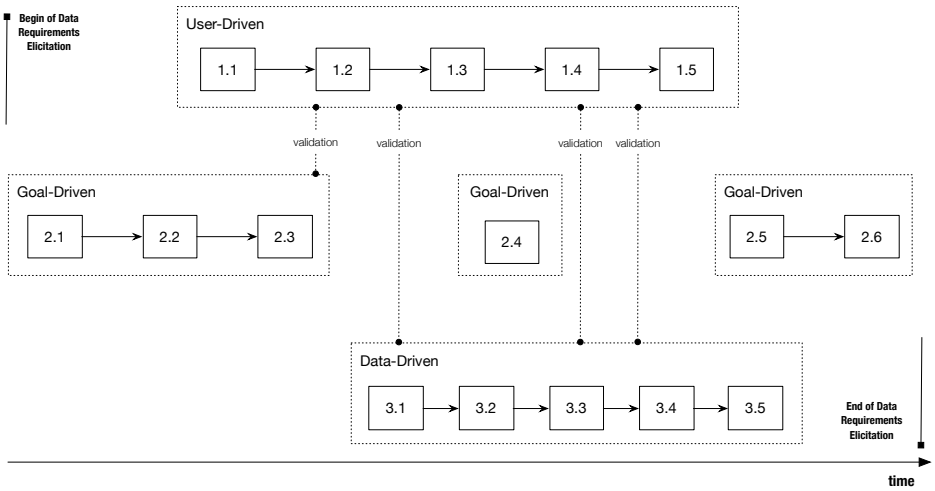

Fig. 1. Methodology for data requirements elicitation in Big Data Warehousing

As the figure suggests, both user-driven and goal-driven approaches were used in parallel, from the beginning of the data requirements definition. This was done by conducting workshops and interviews to: (1.1) obtain knowledge of the organization and problem domain; (1.2) clarification of relevant processes; (1.3) identify relevant data sources and tools to use; (1.4) get users' view of organizational needs; and (1.5) identify existing and new metrics (indicators). It was also important to complement the users' view with the top management's view to: (2.1) align with the organization strategy and business processes; (2.2) identify existing and new metrics and KPI; and, (2.3) gather managers' expectations.

At a given point, the goal and user driven approaches are complemented by the data-driven, to start analyzing the relevant data sources. Throughout this phase, new organizational contexts may arise in which new processes or users must be considered, leading to the need of adding or replacing data sources. These new findings (2.4) must be aligned with the top management and can lead to the identification of new business processes or users. As data is being analyzed, a set of validations is done, as depicted in Fig. 1. In the data-driven approach it is important to: (3.1) map the data sources (which data sources must be considered?); (3.2) conduct the proper data profiling, i.e., describe the data fields; (3.3) classify the available data sources and compare them; (3.4) select and reject data sources or fields (justifying each option); and (3.5) identify the relevant metrics. At the end of the requirements identification, it is important to (2.5) validate the conducted work, aligning all involved stakeholders. Furthermore, it may be necessary to (2.6) prioritize business processes implementation, due to time restrictions, for instance.

4 BDW Prototype Development

Section 4.1 describes part of the developed multidimensional data model, and section 4.2 presents the environment for supporting the decision-making process.

4.1 Dimensional Modelling

For this prototype, 3 fact tables (FT) and 6 dimensions (DIM) were considered, as depicted in Fig. 2. This data is collected from different systems, one of them being SAP. This implied the inclusion of around 100 different attributes, from which, due to confidentiality reasons, only some of them are disclosed in this figure.

The “FT_SpecialFreights” stores records of facts about special freights that were ordered for materials. These are usually expensive and only ordered when material disruptions are imminent. “DIM_SpecialFreight” is a junk dimension of this fact table, integrating different descriptive attributes that give semantics to the stored facts. “FT_Deliveries” stores records of facts related to the arrivals of materials to the plant. These arrivals have a scheduled delivery date and an actual arrival date. This fact table, among other attributes, also stores the difference between these dates, given relevant information about the deviations between the planned and the verified. “FT_EarlyArrivals” stores records of facts concerned with the arrival of materials before the scheduled time. In these situations, the materials are kept on different warehouses than the used for the remaining stored materials. This ensures that suppliers send orders to arrive on the scheduled dates and reduces the warehouse occupation. “DIM_EarlyArrival” is a junk dimension of this fact table, storing a wide range of descriptive attributes. Finally, 4 shared dimensions between the 3 fact tables were considered: DIM_Date, DIM_Time, DIM_Material and DIM_Supplier.

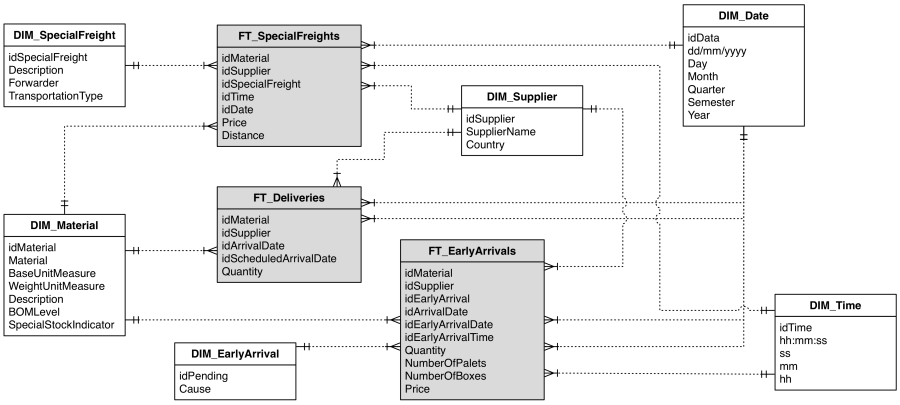

Fig. 2. Multidimensional data model

An evolution of this model is expected. In fact, up to this stage, other 7 data sources were considered and are currently being analyzed: around 250 more attributes were already analyzed and integrated in our data model and it is expected that many more be integrated in the following months.

4.2 Big Data Analytics

For the proof-of-concept of the BDW prototype presented in this paper, the Big Data Analytics component showed in this section only focuses on the total cost of special freights. In this sense, Fig. 3 shows an analytical dashboard created for this analysis.

The map graph shows the total cost of special freights of materials by countries of the respective suppliers. In its turn, at the bottom left, the Treemap depicts the number of special freights from a given supplier and the corresponding total cost of these freights, grouped by country, and using the same color scale of the map graph. The size of the boxes is proportional to the number of special freights, while the color is associated to the total cost of those special freights. At the bottom right, the bar graph gives more detail about the number and values associated to these special freights. Fictitious names were assigned to the suppliers, due to the confidentiality of this data.

As the graphs suggest, the country with more special freights (considering the cost) is Portugal (e.g., supplier s2), with a total cost of more than 600 K €. In its turn, Germany is the country with more suppliers, also presenting a total cost for special frights above the average. On another perspective, Netherlands and Hong Kong are examples of countries with total costs for special freights below the average.

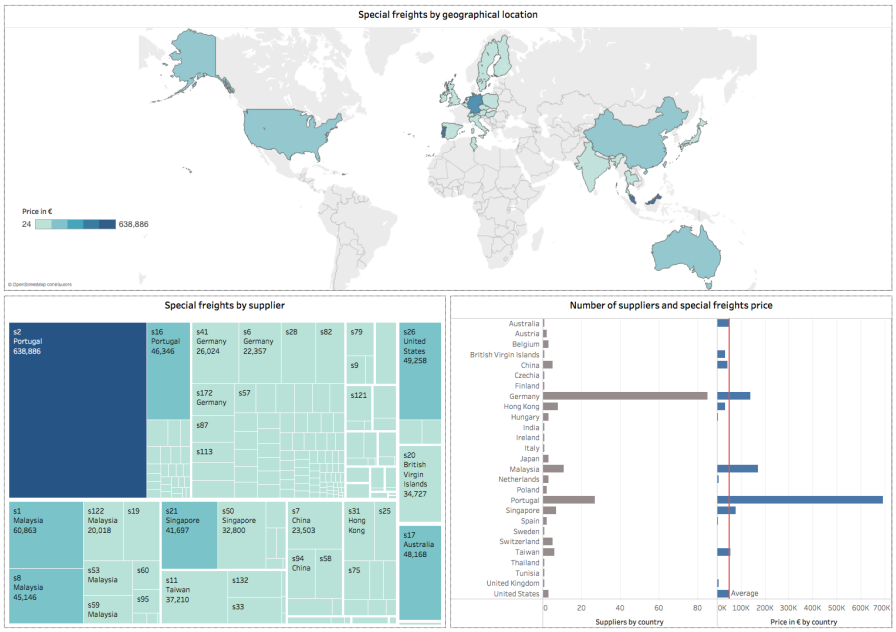

Fig. 3. Interactive dashboard for analyzing data related with special freights

Despite the conclusions withdrawn from the analysis of this data, the purpose of this section is to demonstrate the usefulness and potential benefits of using Big Data Analytics in logistic contexts.

6 Conclusions

Companies that are part of dynamic SCs face themselves with the need of tools to aid in the decision-making process [1]. Most of these solutions respond to particular problems, producing huge amounts of operational data, generated in various formats at increasingly higher rates. Yet, there is still a lack of solutions to integrate this data, contributing to a better decision-making process. The purpose of this paper is to propose a methodology for the data requirements elicitation of a BDW, integrating data from different sources of a company that is part of an international organization of the automotive industry.

The development of the BDW started with the requirements elicitation phase, consisting in applying combined user, goal and data-driven approaches, to obtain the relevant attributes, data sources and KPIs. Despite not being necessary to create a multidimensional data model to develop a BDW, the authors chose this path, because: (1) it allowed a better understanding of the data, organizational processes and relevant KPI to include in the BDW; (2) it ensured the inclusion of the all relevant data, making sure that no important attributes were excluded; (3) it helped in the definition of Hive tables to use. The proposed proof-of-concept makes available interactive dashboards, contributing to a better analysis of the stored data. Next steps include the integration of more data sources. Ultimately, it is expected that the company will be able to perform data analytics methods and, thus, have a proactive approach, rather than a reactive one, when extracting knowledge from its Big Data sets.

Acknowledgements. This work is supported by COMPETE: POCI-01-0145- FEDER-007043 and FCT – *Fundação para a Ciência e Tecnologia* within the Project Scope: UID/CEC/00319/2013; by European Structural and Investment Funds in the FEDER component, through the Operational Competitiveness and Internationalization Programme (COMPETE 2020) [Project no 002814; Funding Reference: POCI-01-

0247-FEDER-002814] and by the Doctoral scholarship PDE/BDE/114566/2016 funded by FCT, the Portuguese Ministry of Science, Technology and Higher Education, through national funds, and co-financed by the European Social Fund (ESF) through the Operational Programme for Human Capital (POCH).

References

1 Levi, D.S., Kaminsky, P. and Levi, E.S.: Designing and managing the supply chain: Concepts, strategies, and case studies. McGraw-Hill (2003)
2 Santos, M.Y., Sá, J.O., Andrade, C., Lima, F.V., Costa, E., Costa, C., Martinho, B. and Galvão, J.: A Big Data system supporting Bosch Braga Industry 4.0 strategy. In: International Journal Information Management, vol. 37, n. 6, pp. 750–760 (2017)
3 Ponis, S.T. and Ntalla, A.C.: Supply chain risk management frameworks and models: a review. In: International J. of Supply Chain Management, vol. 5, n. 4, pp. 1–11 (2016)
4 Kache, F. and Seuring, S.: Challenges and opportunities of digital information at the intersection of Big Data Analytics and supply chain management. In: International Journal of Operations & Production Management, vol. 37, n. 1, pp. 10–36 (2017)
5 Tiwari, S., Wee, H. and Daryanto, Y.: Big data analytics in supply chain management between 2010 and 2016: Insights to industries. In: Computers & Industrial Engineering, vol. 115, pp. 319–330 (2018)
6 Sanders, N.R.: How to use big data to drive your supply chain. In: California Management Review, vol. 58, n. 3, pp. 26–48 (2016)
7 Zhong, R.Y., Newman, S.T., Huang, G.Q. and Lan, S.: Big Data for supply chain management in the service and manufacturing sectors: Challenges, opportunities, and future perspectives. In: Computers & Industrial Engineering, vol. 101, pp. 572–591 (2016)
8 Chen, D.Q., Preston, D.S. and Swink, M.: How the use of big data analytics affects value creation in supply chain management. In: Journal of Management Information Systems, vol. 32, n. 4, pp. 4–39 (2015)
9 Ivanov, D.: Simulation-based single vs. dual sourcing analysis in the supply chain with consideration of capacity disruptions, big data and demand patterns. In: International Journal of Integrated Supply Management, vol. 11, n. 1, pp. 24–43 (2017)
10 Kimball, R.: The data warehouse toolkit: practical techniques for building dimensional data warehouse. In: NY John Willey Sons, vol. 248, n. 4 (1996)
11 Santos, M.Y. and Costa, C.: Data warehousing in big data: from multidimensional to tabular data models. In: Proceedings of the Ninth International C* Conference on Computer Science & Software Engineering, pp. 51–60 (2016)
12 Costa, E., Costa, C. and Santos, M.Y.: Efficient Big Data Modelling and Organization for Hadoop Hive-Based Data Warehouses. In: European, Mediterranean, and Middle Eastern Conference on Information Systems, pp. 3–16 (2017)
13 Santos, M.Y. and Costa, C.: Data models in NoSQL databases for big data contexts. In: International Conference on Data Mining and Big Data, pp. 475–485 (2016)
14 Inmon, W.H.: Building the data warehouse. John wiley & sons (2005)
15 Golfarelli, M.: From User Requirements to Conceptual Design in Data Warehouse Design. IGI Glob. (2010)
16 Abai, N.H.Z., Yahaya, J.H. and Deraman, A.: User requirement analysis in data warehouse design: a review. In: Procedia Technology, vol. 11, pp. 801–806 (2013)