On the Use of Simulation as a Big Data Semantic Validator for Supply Chain Management

ABSTRACT

Simulation stands out as an appropriate method for the Supply Chain Management (SCM) field. Nevertheless, to produce accurate simulations of Supply Chains (SCs), several business processes must be considered. Thus, when using real data in these simulation models, Big Data concepts and technologies become necessary, as the involved data sources generate data at increasing volume, velocity and variety, in what is known as a Big Data context. While developing such solution, several data issues were found, with simulation proving to be more efficient than traditional data profiling techniques in identifying them. Thus, this paper proposes the use of simulation as a semantic validator of the data, proposed a classification for such issues and quantified their impact in the volume of data used in the final achieved solution. This paper concluded that, while SC simulations using Big Data concepts and technologies are within the grasp of organizations, their data models still require considerable improvements, in order to produce perfect mimics of their SCs. In fact, it was also found that simulation can help in identifying and bypassing some of these issues.

Keywords: Simulation, Big Data, Data issues, Semantic Validation, Supply Chain Management, Industry 4.0.

1 INTRODUCTION

Supply Chains (SCs) are complex and dynamics networks, comprised of entities such as suppliers and customers, wherein materials and information exchanges occur, driven by demand and supply interactions. Several activities take place in these networks, such as manufacturing, warehousing and transportation. Ultimately, organizations in these networks aim to fulfil customers' orders at a minimum cost, whilst improving their competitiveness [1]. In other words, to efficiently manage the logistics of the SC, in order to create value for all the agents of the SC.

Simulation is, in fact, among the most used methods for Supply Chain Management (SCM), as postulated by Jahangirian et al. [2] and Pires et al. [3]. Its benefits include the ability to test disruption scenarios, prediction, determine complex solutions, visualization of logistics flows and it can even be combined with other methods, e.g., analytical. However, as the cited authors suggest, its combined use with structures, such as Data Warehouses (DWs), which store, integrate and provide real data to simulation models has not been extensively explored, despite its recognized benefits, e.g., in increasing stakeholders' interest in the solution. The main difficulty to achieve this may be related to the multiple elements that need to be considered in a SC simulation solution. In fact, as suggested by Levi et al. [1], to ensure the efficiency of SCM, multiple activities that take place at a given plant need to be contemplated, e.g., transportation, production and its customers' orders, as well as other aspects concerned with suppliers of suppliers and customers of customers. Only this way will an integrated and holistic data view of the entire network be obtained, hence allowing an efficient SCM.

However, using data of all the relevant business processes of a SC requires vast amounts of data, portraying an environment in which data is generated at increasing volume, velocity and variety, in what is known as the 3 V's of Big Data. Notwithstanding, to the best of the authors' knowledge, no other study has used Big Data concepts and

technologies to store, integrate and provide real data to SC simulation models, despite the benefits that such solutions are expected to bring to SCM, in accordance with Industry 4.0 and as discussed by several studies [4]–[8].

A SC simulation model and a Big Data structure, a Big Data Warehouse (BDW) [9], are currently being developed at a plant of the Bosch organization, using real data from the automotive electronics industry. The BDW stores, integrates and provides data to a SC simulation model developed in SIMIO [10]–[12]. Thereafter, the simulation model can be used to enhance SCM. Notwithstanding, while working on this project, simulation helped to identify several data issues that were not identified with traditional data profiling techniques, which serves as the main motivation for this research. In other words, while traditional data profiling techniques are important to identify syntactic errors in the data that need to be corrected, simulation allowed to uncover other types of issues, which are henceforth referred to as semantic issues.

In light of the above, the purpose of this paper is threefold. First, it describes how simulation was used as a semantic validator of the data that was used, hence leveraging typically used data profiling techniques. Second and using such insights provided by simulation, this paper proposes a classification for the data issues that simulation helped to identify, while working on the SC simulation model in a Big Data context. Third, this paper also quantifies and provides a discussion of the impact that such issues had on the volume of data that was used in the final instance of the SC simulation model. This way, the authors hope that future researches in similar domains find the insights and conclusions shared in this paper to be useful, when developing similar projects and facing equivalent difficulties.

This paper is organized as follows. Next section analyzes and discusses literature related to this research. Section 3 describes the methodology for this research, the Big Data concepts and technologies that were applied and the automotive electronics SC that is being considered. Section 4 addresses the issues that simulation allowed to uncover in a Big Data context. In light of this, a classification for such types of issues is proposed and its impact in the final simulation solution is analyzed. Section 5 discusses and analyzes the main insights that can be withdrawn from a managerial perspective. Finally, conclusions and future research directions are provided in the last section.

2 RELATED WORK

This section discusses and analyzes literature related to the research addressed in this paper. Thus, first subsection analyzes simulation studies that were combined with structures to store and integrate data in several domains. Next, second subsection discusses the benefits that are expected from the combined use of Big Data and simulation in SCM. Finally, last subsection provides some concepts related to Big Data Warehousing.

2.1 Simulation and Structures to Store and Integrate Data

Bottani [13] reported the use of a Simul8 model representing logistic movements in a warehouse. It allowed to observe the impact of RFID in a warehouse, including the data that would be generated by such identification tags e.g., expiry date, type of component, production lot. Thereafter, this data was collected and stored in a DW, which, in its turn, was used to extract information from the simulation results. The authors emphasized the use of the DW structure to derive value-added information from the simulation results.

Gupta [14] proposed an ExtendSim simulation model that allowed to simulate different DW query workloads to meet performance objectives. According to the authors, the workload was managed considering query scheduling,

admitting and executing, while also allocating resources to meet the performance objectives. This way, the authors could experiment new workload schedulers without trying them on the real DW.

Advocating that typically used commercial tools present simulation results as averages that mask important aspects of the transient behavior of the system, Ehmke et al. [15] developed a DW that stored and allowed analytics to be performed on the obtained simulation results. This way, it was possible to extract information from the simulation results by aggregating this data. The simulation model was developed in Plant Simulation and the authors showed the benefits of such approach in a real case study, which considered transportation scheduling, operational procedures and infrastructural changes in the Mississippi River waterway system.

Postulating that effective data cleaning approaches produce significant savings, since bad decisions can be taken based on bad data cleaning, Li and Joshi [16] proposed a simulation model to help practitioners determine the best data cleaning strategies to use. The authors used ProModel to test 2 different data cleaning approaches, which were executed during the data integration phase, i.e., the usual extract, transform and load of data into the DW. According to the authors, simulation allowed them to better understand the interactions among data cleaning approaches and the produced results, which could not have been just as effectively achieved with other methods.

Nageshwaraniyer et al. [17] proposed an Arena simulation model of material handling operations of a coal mine real case study. In this study, the DW is used to feed data to the simulation model, while the results obtained by the later are also sent to the DW for analysis and further reporting. According to the authors, the use of both technologies allowed the best parameters for the system to be determined.

As postulated by Kugu et al. [18], data mining methods allow information to be extracted from data sets. The authors developed a simulation model, which aimed to, using intelligent learning capabilities, generate data for the situations in which there is not enough data stored in the DW to allow such patterns to be discovered. This way, it was possible to apply the data mining methods, despite the lack of real data stored in the DW. The authors concluded their study by emphasizing the benefits originated from their approach.

Truong and his coauthors [19]–[21] proposed a conceptual framework which can be used for simulation models dealing with high volumes of data. The simulation model uses data stored in a DW, combining both real data and simulation data from previously conducted simulation experiment. In its turn, the analytics part of the framework allows typical data aggregation and analysis of data produced by the simulation model. According to the authors, their framework would not be suitable for simulation models dealing with small amounts of data, due to the considerable time required to implement such framework. Truong et al. [20] used this framework and proposed a way to automatically calibrate the simulation models, i.e., the process of tuning model parameters, so that the reliability of the simulation model increases.

Rabe and Dross [22] proposed a framework comprised of a DW and a simulation model. The main purpose was to use simulation to predict changes that would occur to specific changes in the logistics network of a company. In this framework, a DW measured Key Performance Indicators (KPIs) to allow analytics, while a copy of such DW was created to store simulation results. These simulation results, in their turn, were used as reward criteria for reinforcement learning algorithms, employed in the simulation model.

As the above reviewed studies suggest, the scope of most papers is reduced to a specific process, e.g. warehousing, not considering all the activities occurring in a SC. Furthermore, all the reviewed papers considered traditional databases or DWs, not using the benefits of Big Data concepts. Table 1 summarizes these findings.

Table 1: Summary of the extracted information from the analyzed studies that combined simulation with structures to store and integrate data.

Author(s) (year)	Simulation Tool	Simulation and DW usage
Bottani [13]	Simul8	• Analytics on the simulation results
Gupta et al. [14]	ExtendSim	• Simulation of different workload schedulers for a DW
Ehmke et al. [15]	Plant Simulation	• Analytics on the simulation results
Li and Joshi [16]	ProModel	• Simulate different data cleaning strategies for a DW
Nageshwaraniyer et al. [17]	Arena	• DW feeds the simulation model with data • Simulation results sent to the DW
Kugu et al. [18]	NA	• Simulation used to generate data for data mining algorithms
Truong et al. [19], Truong et al. [20] and Truong et al. [21]	NA	• DW feeds the simulation model with data • Simulation results sent to the DW
Rabe and Dross [22]	SimChain	• DW feeds the simulation model with data • Simulation results sent to the DW

These results show the usage that is given to the combined use of simulation and the data integration capabilities of structures such as DWs. From these, 3 main usages can be highlighted: (1) use analytics to extract additional information from simulation results stored in a DWs; (2) use DWs to feed data to the simulation model; and (3) use the DWs to integrate simulation results. Apart from these, other usages were also identified, namely: use simulation to test different workload schedules for a DW or different data cleaning strategies and use simulation to generate data to apply data mining algorithms. This shows that, while few studies have done so, simulation has been used to improve certain phases of a DW lifecycle. In fact, the approach proposed in this paper uses simulation to improve one of the development phases of the BDW, as will be described in section 3.

Finally, few of the analyzed studies considered SCM problems. In fact, Jahangirian et al. [2] also suggested the lack of studies that combined the use of simulation and structures to store and integrate real data. In light of this, next subsection reviews studies that highlight the role that Big Data can have in leveraging the quality of SC simulation solutions.

2.2 Simulation and Big Data in SCM

The need to improve industrial processes is, in fact, one of the main goals of Industry 4.0 as is emphasized by Kagermann et al. [7]. Such improvement may involve several methods, with the authors stressing the use of simulation to analyze the behavior of complex systems such as SCs. Simulation is even mentioned in one of the example applications provided by the authors, to analyze risk scenarios in SCs. The authors also noted the importance

of using Big Data concepts and tools with such solutions, as it allows vast volumes of data from several data sources to be considered in the model.

Vieira et al. [5] reviewed simulation studies closely related to the concept of Industry 4.0, in order to identify the boiling research directions for simulation, which are aligned with the Industry 4.0 movement. According to the authors, such studies include the use of Big Data concepts and technologies applied to SC problems, due to the possibility of capturing the detail of processes that Big Data allows, along with the ability to consider the uncertain nature of SC systems that simulation allows.

Zhong et al. [4] outlined the current movements on the application of Big Data for SCM. According to the authors, the increasing volume of data in the several SC sectors is a challenge that requires tools to make full use of the data, with Big Data emerging as a discipline capable of providing solutions for analysis, knowledge extraction, and advanced decision-making. Thus, according to the authors, the ability to use such structures to provide data to simulation models, allowing the latter to include data from several sources relevant for the problem, should be considered by organizations.

According to Tiwari et al. [6], the use of analytics in SCs, including simulation methods, is not new. However, the advent of Big Data presents itself as an opportunity for its combined use with such analytics methods. In particular, the authors stress the importance of simulation and Big Data in predictive and prescriptive analytics, with simulation being used in the former to predict future events and in the later to enhance alternative decision-making testing.

As the cited works suggest, and to the best of the authors' knowledge, a gap can be identified in literature, which consist in the lack of Big Data structures to store and integrate data from several sources, with the end goal of providing such data to a SC simulation model. Thus, next subsection reviews concepts related to Big Data Warehousing.

2.3 Big Data Warehousing

In today's world, data is generated at increasing velocity, volume and variety, in what is known as the 3 V's, or three main characteristics that portray a Big Data environment [23]. In light of this, it becomes harder for traditionally used structures (e.g., DWs) to process such volumes of data [24]. In this regard, different types of solutions have been proposed and implemented. As Costa et al. [25] and Costa and Santos [26] suggest, some considered implementing DWs in NoSQL databases, albeit these solutions only scale the operational systems (see [27] for a comparison of NoSQL engines). Eventually, SQL on-Hadoop emerged as a more efficient solution for Big Data contexts [25], [26], [28], [29]. See [30] for a comparison of Hadoop and other alternative solutions for Big Data contexts and Grover and Kar [28] for a summary of existing Big Data tools, including the Hadoop ecosystem.

Hadoop is an ecosystem based on the MapReduce programming model and the Hadoop Distributed File System (HDFS). This ecosystem includes several tools, from which Hive and Impala can be emphasized, since they will be addressed in this paper, namely in subsection 3.2. Hive was created by Facebook to improve the query capabilities of the Hadoop ecosystem, which were limited and not very productive [31]. In addition, it is extensively used in many organizations for reporting, ad hoc querying and analysis [32]. Hive has its own query language: the HiveQL. Furthermore, it organizes data in tables (each table corresponds to an HDFS directory), partitions (sub-directories of the table directory) and buckets (segments of files in HDFS). In fact, when dealing with performance

issues, these aspects must be carefully analyzed, as discussed in more detail in the work of Costa et al. [33]. Thus, a DW developed in Hive can be seen as a BDW, being a flexible, scalable and highly performant system that uses Big Data techniques and technologies to support mixed and complex analytical workloads, e.g., streaming analysis, ad hoc querying, data visualization, data mining and simulations [9].

Santos et al. [34] presented a Big Data system architecture implemented in Bosch Car Multimedia in Braga, Portugal (the same plant of the case study considered in this paper), which supports the Industry 4.0 technological movement followed by the organization in question. The developed Big Data system integrates data from several business processes (e.g., customer quality claims), allows the analysis of several KPIs and was implemented in Hadoop. Nodarakis et al. [35] extracted hashtags from large scale tweets to classify them into different sentiments, in a parallel and distributed manner, using the same ecosystem. The authors also conducted experimental evaluations to prove their solution is efficient, robust and scalable, therefore being appropriate for Big Data contexts. Kv and Kavya [36] also used Hadoop to analyze trends of e-commerce web traffic logs.

3 METHODOLOGY

This section addresses the methodology followed for his research. In this regard, first subsection describes the SC system that was considered. Next, last subsection describes the approach that was followed in this research, which culminated in obtaining a valid and coherent SC simulation model, i.e., without semantic issues. It is important to address this as it allows the importance of simulation, in validating the semantics of data, to be perceived.

3.1 Supply Chain Characterization

This project is being developed at a plant of the Bosch Group, which produces electronic components for cars. Around 7 000 different types of materials are actively being supplied by roughly 500 different suppliers, located in more than 30 countries, especially from Europe and Asia. Germany, Netherlands, Malasya, Taiwan, China, Hong Kong and Singapore are some of the countries with suppliers actively providing materials to the manufacturing company located in the north of Portugal. In fact, an ordinary car is comprised of multiple different types of materials. Therefore, in this type of industry, materials are often supplied by single sources typically located in different countries, hence exposing the producers to eventual disruptions [37].

Like most automotive electronic plants, this one also follows a demand-driven production approach, aiming to reduce overall wastes with inventory levels and other aspects [37], [38]. However, whilst this may result in such benefits, it also may result in high vulnerabilities [39], if the available materials are not enough to cover eventual disruptions or quality problems. To mitigate these risks, safety stock strategies are usually applied in this type of industry, as is the case with the company in question. In this regard, orders are placed to suppliers and monitored, so that when the right time comes, they can be sent to the plant, in order to arrive at the scheduled date, hence reducing the need to create material buffers. Figure 1 illustrates a summary of the main material and information flows that occur in this system.

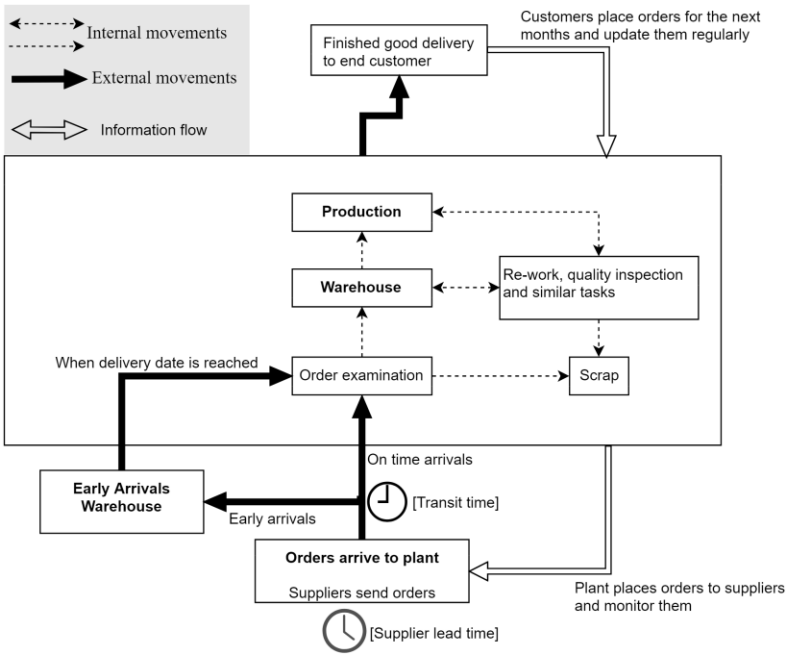

Figure 1: Summary of material and information flows of the SC system.

As Figure 1 illustrates, most of these arrivals occur at the scheduled date, however, some suppliers provide the orders before the scheduled date. When these situations occur, the plant stores such orders in a warehouse dedicated to early arrivals, which is managed by suppliers, so that the plant does not incur in excessive warehousing costs. In its turn, when suppliers are delayed, the plant may schedule special freights, which are considerably costlier, albeit much faster. Thus, whilst early arrivals results in high warehousing costs for suppliers, late arrivals may result in material shortages, potentially leading to production stoppages.

When orders arrive to the plant, their contents are examined to assess their quality. If the quality of the materials is not according to the standards, the materials are scrapped and interaction with the respective supplier is initiated. Notwithstanding, most materials do not have quality problems and, as such, can be stored in the warehouse. When required by production, the materials are consumed, so that the ordered finished goods may be delivered in due time to the respective customers. After having been stored and if requested, materials may also be sent to quality assessment, re-work and other similar tasks, to ensure the quality of the finished good when its production is finished. Such strict quality controls and, at the same time, the high levels of product customization required by increasingly demanding end customers, are normal in this type of industry, as recognized by Masoud and Mason [38] and Simchi-Levi et al. [40].

This section summarized the size of the SC system in analysis and the types of flows that occur in the plant, which are elements that must be considered in a SC simulation. Such description is important, in order to understand the relevance of the data issues that are addressed in section 4. Next subsection describes the approach that was conducted, in order to reach a valid and coherent data model of the SC system at hand.

3.2 Proposed Approach

When using real data in a simulation model, the integration of such data must create a coherent model, i.e., in order to accurately mimic a process, all its elements must be present and coherent, and the respective data cannot have semantic issues. While data profiling techniques are mainly used to evaluate aspects such as errors correction, conflict resolution and null values treatment, the authors argue that these only evaluate the data in a syntactic level. Thus, this section describes the approaches that were adopted to allow simulation to be used as a semantic validator of the data stored in the BDW, so that a coherent simulation model can be obtained. In this regard, to provide an overview of the development of the BDW and the SC simulation model, each of its main development phases are depicted in Figure 2.

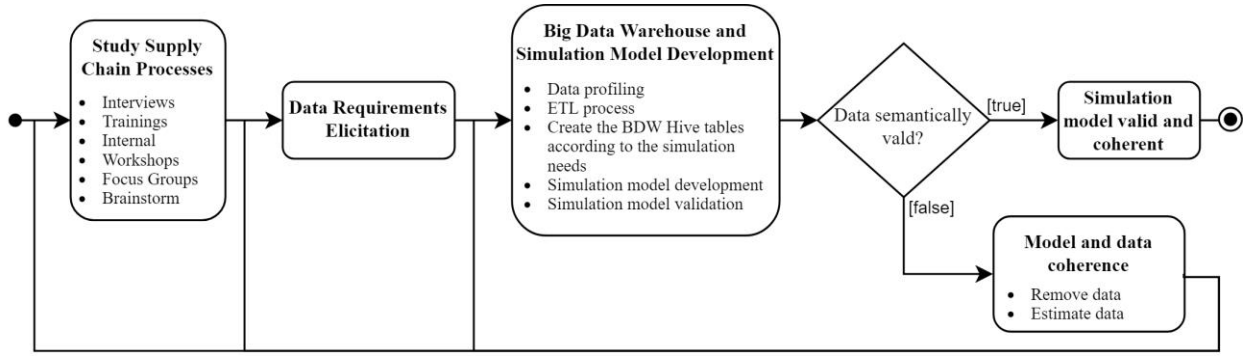

Figure 2: Phases of the project.

SC processes differ among organizations, geographic locations, businesses, industry sectors and other aspects. Thus, it is mandatory to start by studying the processes associated to the SC at hand. For this purpose, interviews, workshops, trainings with process specialists and other types of sessions were helpful to gain insights about relevant SC processes. Next, the data sources used at the plant to manage the relevant business processes need to be analyzed, since these are the ones that produce relevant data for the simulation model. Thus, this phase consists in analyzing them, in order to select the most important variables and start designing the data model of the BDW. In fact, in Big Data contexts, data models are not usually a main concern in terms of providing an overall and integrated view of the data. However, this step is important to be considered at this stage, as it provided the following major benefits, which are discussed in more detail in [41]: better understanding of the data sources, organizational processes and relevant variables to include in the BDW and in the simulation model; making sure that no important data is excluded; and helped in the definition of the BDW model, namely the Hive tables to use.

Having selected the relevant variables to use in the project, it is necessary to assess the quality of the data, in what is known as data profiling. In this stage typical errors, e.g. null values, format evaluation and other syntactic errors are detected and documented, to be later corrected. Once this evaluation is finished, the development of the BDW and the simulation model can be initiated. Such development is conducted in successive iterations, which may even require other data sources to be analyzed and possibly included, in successive iterations. Thus, as depicted in Figure 2, this process is conducted until the model is considered to be valid and using semantically validated data, i.e.,

until a coherent simulation model. In this regard, Figure 3 shows the steps that were performed in order to provide semantically valid data to the SC simulation model.

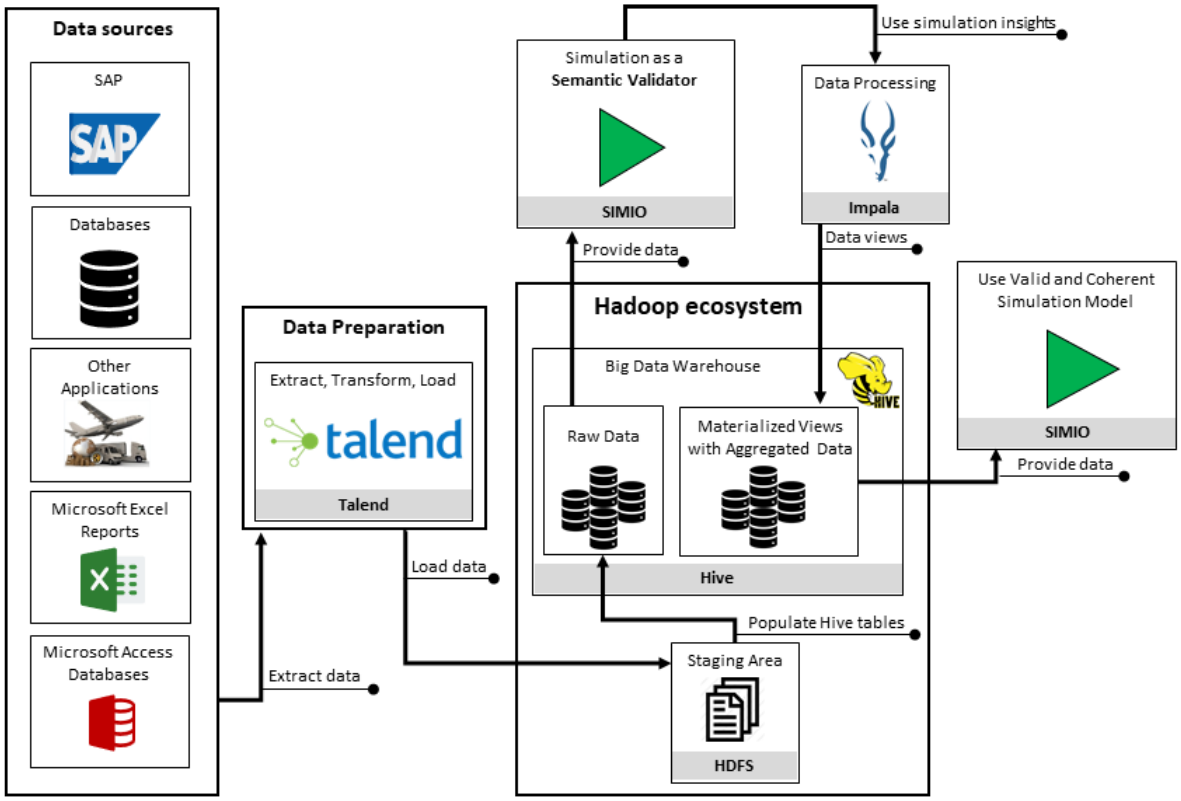

Figure 3: Steps required in order to provide semantically valid data to the simulation model.

As can be seen, the first step consists in the typical ETL process. In it, data from selected data sources is extracted, transformed and loaded to the HDFS, with the data transformations that occur in this process being responsible for correcting the syntactic errors that were identified in the data profiling phase. Hence, the data that is sent to HDFS is clean and does not contain any syntactic errors.

Afterwards and using the insights obtained from the study of SC processes and its respective data sources, the data in the HDFS can be modelled according to the simulation requirements and used to populate the associated Hive tables. Since the Hive tables are modelled according to the simulation needs, it is important to, at the same time, develop the SC simulation model. This way, there is no need to search for any value during simulation runtime, which, in Big Data contexts, should be avoided, otherwise the simulation run may become considerably slow.

At this stage, real data from several sources is integrated and stored in the BDW and can be provided to the simulation model that has also been validated. See [42], [43] for references to simulation model validation techniques. In addition, such data does not contain syntactic errors. However, when providing it to the simulation model and using it to try to reproduce a copy of the real system, additional issues are detected, which were not detected in the data profiling phase. Such issues are discussed in section 4.

Using the insights provided by simulation, data views on the Hive tables are used to create additional ones that reflect the corrections that need to be made. These corrections may involve exclusion of certain records, aggregations of values and other operations. Such data views are performed using Hive Query Language (HiveQL). In fact, this correspond to a partial materialization of an OLAP (Online Analytical Processing) cube, as is discussed in more detail in the work of Correia et al. [44]. In addition, apart from using real data, there may also be the need to estimate data, e.g., using statistical distributions. All these considerations are achieved by using simulation as a semantic validator of data. Finally, with the new set of Hive tables that store semantically validated data, the simulation model can use it, hence allowing coherent simulations to be conducted.

4 ANALYSIS OF DATA ISSUES FOUND IN A BIG DATA CONTEXT

While developing a SC simulation model in a Big Data context, several data issues were faced, which needed to be handled, in order for the simulation model to maintain its coherence. Some of these issues could be identified with typical data profiling techniques, while others could only be identified and quantified due to use of simulation as a semantic validator of data. Thus, in the first subsection of this section, a classification for the latter is provided. Next, the impact of such issues is quantified in the last subsection.

4.1 Data Issues Classification

The aim of this subsection is to propose a classification of the data issues that simulation helped to identify while developing a SC simulation model in a Big Data context. The following is a list of such categories of issues.

- **Data not according to a business process (A)**

This category comprises the cases in which an analysis of the data shows that a given business process is not represented by the data. In this work, such example consisted in the case of the internal material movements, the analysis of which showed that the storage strategy followed by the plant's warehouse was not represented in the data. This is especially interesting because such movements are stored in SAP (Systems, Applications and Products) and, supposedly, all movements must be registered. In addition, this is an example of a data issue that was only discovered when conducting the first simulations using this data, meaning that syntactic errors were corrected in the data profiling phase, however, when the data was used in the simulation model, it was possible to observe that the behavior of the real system was not mimicked by the simulation model. Thus, if simulation was not used in this project, this issue would hardly be detected. For instance, if the BDW was used to provide or feeding data to machine learning algorithms, these would learn from such data, potentially leading to wrong generalizations [45].

Since simulation was used to identify this issue, it can also be used to quantify it. In this regard, Figure 4 shows the percentage of movements that occurred but did not represent the storage strategy followed at the plant, differentiated by movements to and out of the warehouse.

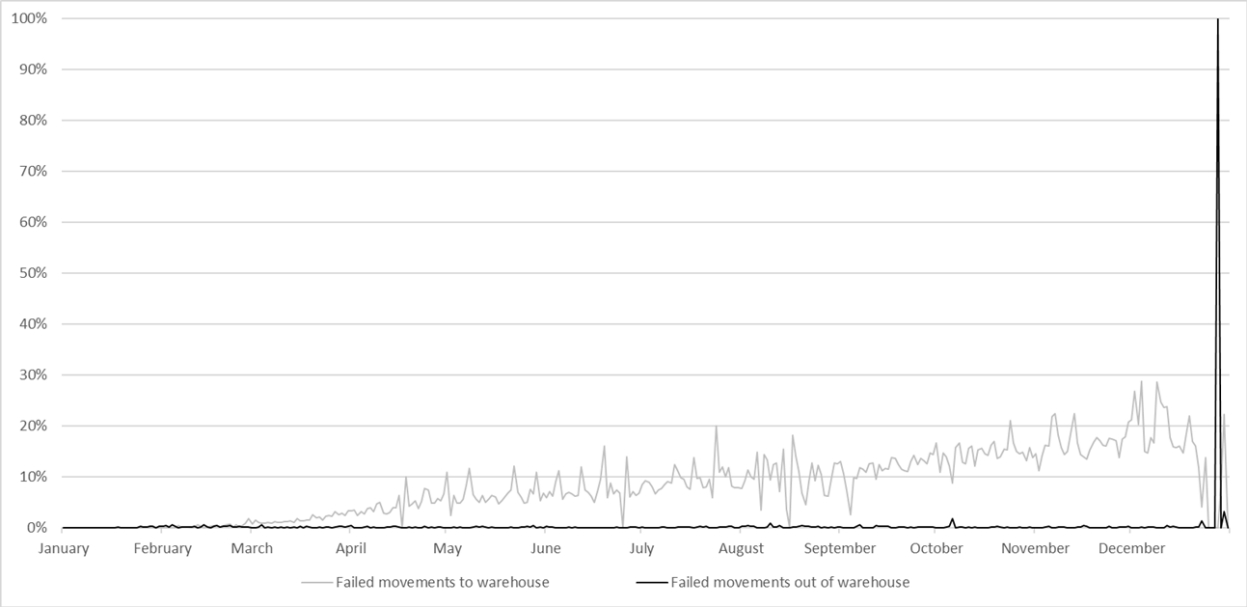

Figure 4: Percentage of material movements not according to the storage strategy followed at the plant.

As can be seen, the number of movements that do not represent the strategy followed at the plant is considerable and happens throughout the entire year. In this regard, when discussing this issue with managers, some explanations could be found, e.g., not all movements are registered in SAP. Nevertheless, it was not possible to discover the real cause for this problem. Therefore, the true implications of this issue could not be determined, e.g., if all movements are registered, then movements may be registered with a wrong date. In fact, in this particular case, this issue lead to the need of simplification of the problem and redesigning the data structure that models the warehouse in the simulation model. However, if the original approach had not been taken, this issue would not have been discovered, as per the approach described in Figure 3, subsection 3.2.

- **Wrong data (B)**

This category comprises the cases in which data does not have any syntactic problem, however, when modeling the data according to the simulation needs (as discussed in subsection 3.2) such errors are identified. Examples found in this project include: order dates prior to the associated arrival, and sum of an order's transit time with the order date is higher than the associated arrival date.

- **Missing data (C)**

This category comprises the cases in which a given data source could not be obtained. The main causes that lead to this issue are: non-existing data sources, lack of standardized data sources, no historical data and no access to data. These are next discussed.

The former, as the name implies, suggests that the data source for a given business process does not exist in the organization. Examples for this issue are the data concerned with production times and capacity and data related to suppliers, such as their capacity and their multi-tiers' suppliers.

The lack of standardized tools portrays an interesting challenge, as the data, in fact, exists, however, managers use different data sources, which potentializes the risk of entering contradictory data in the simulation model. In this

project, such case was found with the Bill of Materials (BOM). As discussed in subsection 3.1, the environment considered in this research is characterized by having thousands of raw materials, which have usage in hundreds of finished goods and in which a single raw material may have usage in multiple finished goods. Furthermore, the BOM of products may change with time. Thus, the production is divided in multiple department subunits. With different Information Systems, Excel files and others being used to manage this data, contradictory data may be stored among the different department subunits. Depending on the situation, these issues may be solved with meeting with process and data experts. However, in more complex situations, this problem may lead to the exclusion of the data sources.

Historical data allows facts that occurred at a given period of time in the past to be obtained, adding calendar context to that business indicator. Examples of this issue include: historical stock level and warehouse contents; historical materials' master data; and historical suppliers' data, such as travel mode and transportation duration. Thus, without this data, simulation cannot replicate the facts that occurred in the real system as the simulation clock advances. For instance, without knowing the historical contents of the warehouse, it is not possible to set the stock at the beginning of the simulation. In this regard, other approaches need to be considered, e.g.: test different safety stock calculation methods. See Schmidt et al. [46] and the therein cited references for a review of safety stock calculation methods. In fact, simulation can be combined with analytical approaches to test such adequate levels of safety stock to use, when historical stock data is not available. This way, it would be possible to estimate the contents of the warehouse at the beginning of the simulation.

Lastly, working in simulation of SCs in Big Data contexts typically entails working with data that is sectorized in different departments, making the access to such data difficult. Being sensitive data, even with the interest and involvement of stakeholders in the project, organizations may be reluctant to share the data. In this case, such was the case with data concerned with final customers' orders, forecast of final customers' orders and production scheduling.

The need for systems thinking, which simulation practitioners must employ, in order to understand what business processes must be present in a given simulation model, may help in identifying the data requirements that otherwise would not be easily identified. Thus, when it is determined that a given business process is relevant for simulation, albeit there is no real data to model it, other approaches must be considered.

- **Incomplete data (D)**

This category comprises two cases. The first is concerned with the cases in which the data attributes exist, the data values are not null, however, such values do not provide the information that is required by simulation. For instance, in the case of the date of the special freights, whilst its date attribute exists, it is used to store the intended date for its arrival and not the actual arriving date, leaving the doubt if the freight actually arrived at the desired date. I.e., the level of information entered in this data source is not enough to produce a reliable simulation. Finally, all the data sources used to store data related to arrivals of materials to the plant allow manual inputs of materials and their respective quantities, opening the possibility of users entering values that cannot be treated in the ETL process, with users also applying different terminologies to specify multiple arrivals in the same field. The mentioned examples provided by this issue create a problem related to the arrival of materials to the plant, as the data of the multiple data sources provides incomplete information.

The second case is concerned with the situations in which it is found that data is missing. Contrarily to the first type of issue, in this one it is known that data is missing. For instance, when integrating the data related to orders to suppliers and the data related to material arrivals to the plant, as per the approach described in subsection 3.2, it was found that there were arrivals that were not ordered, as well as orders that did not arrive.

- **Data not maintained (E)**

This issue comprises the cases in which data exists, the values are not null, albeit they are not maintained. As such, the values stored in these attributes are not accurate. For instance, despite SAP providing fields to insert material's shelf life or scrap rate, it was found that users insert standard values in certain fields, e.g. a given value for a plant and a different one for another plant.

In a Big Data context with hundreds of variables being analyzed, such issues may only be identified when using the simulation as a semantic validator of the data (see Figure 3 and the approach described in subsection 3.2). In fact, as discussed in [41], in Big Data environments, in which real industrial data is being analyzed, it is common for different views and understandings of the data to exist. Hence, simulation may help in detecting these cases, as incorporating them in the model results in incoherent simulations.

- **Data conflicts (F)**

Some data sources have attributes in common, leading to situations in which even managers contradict themselves in determining which should be used. It is therefore mandatory to meet with process and data experts to understand which attributes should be used. Like the previous issue, in this one, simulation may also help in detecting these cases, as incorporating incoherent data in the model results in equally incoherent simulations. Examples of this issue consist in the supplier' countries and their orders' transit time, as all data related to suppliers is managed by an Excel file and later inserted in SAP, creating cases in which both data sources have relevant attributes, albeit, only one should be used, otherwise data inconsistencies can occur.

In the above discussion, several business processes were provided as examples for the data issues being discussed. While the mentioned examples may not be mandatory to include in a simulation model that simply aims to mimic the behavior of the real system, if the purpose is to leverage the benefits of simulation, e.g. to test disruption scenarios, then, the mentioned data should be considered, in order to achieve efficient and accurate representations of the real system.

Being among the pioneers that applied Big Data concepts and technologies to provide real data to a simulation model, the authors proposed the above classification of the issues that were faced, aiming to provide a common ground for future researches of SC simulation in Big Data contexts. Table 1 lists the issues that were faced in the project and classifies them according to the proposed classification.

Table 2 Summary of the identified data issues and their classification.

	Data issue category					
Data issues	A	B	C	D	E	F
Arrival date prior to order date		■				
Date of special freights				■		

Data issues	Data issue category					
	A	B	C	D	E	F
Historical data (e.g., stock, material's price)			■			
Lack of BOM			■			
Lack of customers' orders			■			
Lack of production capacity			■			
Lack of production plans			■			
Lack of production time data			■			
Lack of suppliers' data (e.g., lead time, capacity, multi-tier suppliers)			■			
Lack of forecast			■			
Materials quantity in different data sources used to manage material arrivals				■		
Materials reference in different data sources used to manage material arrivals				■		
Materials' ABC classification in different data sources						■
Materials' dimensions and volume					■	
Materials' safety time in different data sources						■
Materials' shelf life					■	
Nonexistent orders' registers				■		
Date of order to supplier+ transportation duration > order's arrival date		■				
Scheduled delivery data of orders to suppliers in different data sources						■
Plant's and suppliers' scrap rate of materials					■	
Suppliers' city					■	
Suppliers' country in different data sources						■
Suppliers' geographic location			■			
Storage strategy specified by the data	■					
Suppliers' transportation duration in different data sources						■
Suppliers' transportation duration					■	
See the meaning of the column names in the above list						

Indeed, the subject of data quality problems, data issues, or dirty data is not new [47], [48], not even for the simulation community [49]. In fact, Bokrantz et al. [49] presented a multiple-case study within the automotive industry to provide empirical descriptions of data quality problems in simulation projects. As the authors postulated, simulation requires high-quality data and, often, extensible transformations to allow its utilization in simulation models, i.e., data issues must be bypassed, in order to produce a coherent simulation model. In its turn, on the note that there is no widely adopted classification for data issues, Laranjeiro et al. [48] surveyed the literature concerning such problems and mapped them according to the dimensions they identified.

In fact, the types of data issues that exist in a Big Data context are expected to be the same as the ones that exist in traditional environments. However, the classification of data issues provided in this research is concerned with the ones that simulation helped to identify in a Big Data context and inserted in a real industrial environment. Such an example is provided by the issue of missing data, as such gap was identified with the help of simulation, namely by its need for systems thinking; typically, such problem is not classified as a data issue, as the data, in fact, does not exist. In light of this and since this research, to the best of the authors' knowledge, is among the pioneers that developed

a SC simulation model combined with Big Data concepts and technologies, a fortiori, this is also the first study that proposes a classification for such issues. Next subsection provides a discussion of the impact that such issues had in the volume of data provided to the final version of the simulation model, i.e., the volume of data that is semantically valid.

4.2 Quantitative Analysis of the Data Issues

In this subsection, the data issues found in this project are quantified and its implications to the volume of data used in the coherent simulation model are also analyzed. In this regard, Figure 5 shows the number of attributes, data rows and volume of data extracted from the considered data sources, stored in the BDW as a first load from the ETL process and materialized using the data views, after using simulation as a semantic validator, as per the approach described in Figure 3, in subsection 3.2.

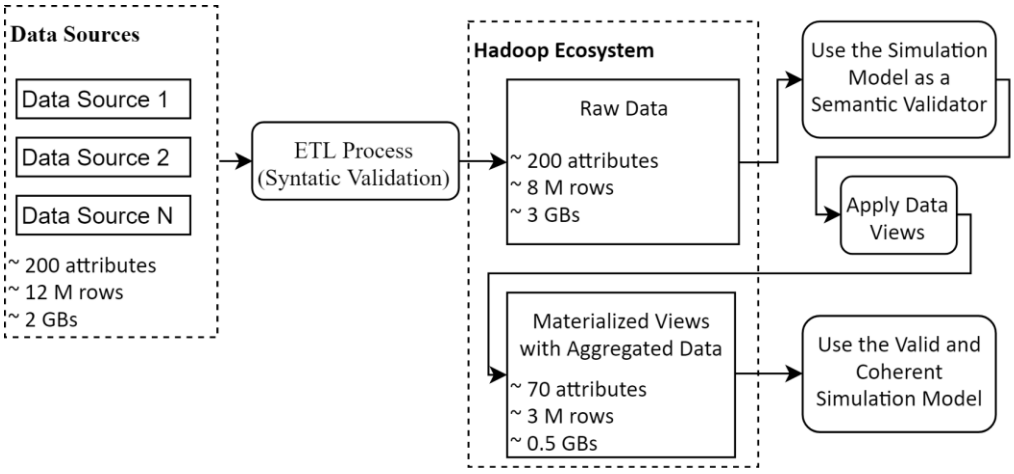

Figure 5: Analysis of the impact of the data issues in terms of number of attributes, data rows and volume of data in GBs.

During the process of analyzing the SC processes and the associated data sources (see Figure 2 in subsection 3.2), around 2 000 attributes were analyzed. From those, as the figure suggests, around 200 attributes were considered in the ETL process and loaded to the BDW (Raw Data in Figure 3 and Figure 5). In this regard, despite having extracted around 12 million of rows from the data sources - roughly 2 GBs of data – the ETL process culminated in storing 8 million rows – roughly 3 GBs of data – in the BDW. This data corresponds to a year of activities in the SC at hand. This occurred, because some rows are filtered in the ETL process, due to syntactic errors, e.g., null dates or untreatable values. Afterwards, using Big Data concepts, such as fully denormalizing the data so that there is no need to search any value during the simulation run, results in considerably increasing the volume of data.

Next, the data, supposedly without any issues, is provided to the simulation model. However, its utilization reveals additional issues that need to be treated, otherwise the produced simulations are not reliable and coherent. I.e., using the insights obtained by using the simulation model as a semantic validator of the data, data views are executed, in order to create additional Hive tables. Thus, the obtained data can be provided to the simulation model since it was semantically validated. This resulted in the exclusion of some data rows and some attributes, with some examples for the reasons of this exclusion being provided in subsection 4.1.

As described in subsection 3.2, the process of providing semantically valid data to the simulation model may include not only removing data, but also estimate data and even the use of statistical distributions to model specific business processes. For instance, if no production times are available, statistical distributions must be used, so that the object that models the production activities of the plant remains busy during such task. Because of this, despite being provided with semantically valid data, this does not mean that the model will produce exact mimics of the real system. Notwithstanding, without such semantically valid data, the model would produce inaccurate simulations. Hence, despite the high volumes of data that are already available, simulation experts will still require to use traditional approaches, e.g., distribution fitting, in order to be able to use the real data in their models.

These results also show that, on one hand, despite having analyzed considerable volumes of data, the final simulation model that is obtained used a reduced part of such data. On the other hand, data for certain business processes could not be obtained, which would considerably increase the volume of data. Hence, it is, indeed, arguable if this environment can be considered a Big Data one. Notwithstanding, as Madden [24] and Kaisler et al. [50] stated, there is no widely accepted threshold value for a Big Data context. Thus, often, this value is defined as the volume that exceeds the capacity of traditional tools to process the data, hence being necessary to apply Big Data concepts and technologies. Such was the case with this research, where the organizations' Big Data cluster was used and Big Data concepts (namely by denormalizing the data) and technologies (namely the Hadoop ecosystem and the therein included Big Data tools, e.g., Hive and Impala) were applied. Furthermore, this research considered data of 1 year. Thus, if the volume of data keeps increasing, as per a Big Data context, and if the mentioned data issues are solved, the need for the employed Big Data concepts and structures becomes even more significative.

5 DISCUSSION AND MANAGERIAL IMPLICATIONS

One of the main implications provided by this research is that, in a Big Data context, simulation was found to be an essential tool to detect, quantify and understand some of the issues. Notwithstanding, the ETL process is still vital, as it allows certain syntactic errors, identified in the traditional data profiling phase, to be corrected. However, simulation, takes this verification to a different level of exigency, since there is an obligation to integrate data, in such way that it must originate a coherent model. I.e., in order to accurately mimic a process, all its elements must be present and have no semantic issues. In this regard, the authors hope that managers find the contributes and experiences shared in this paper useful when developing simulation projects in Big Data contexts.

This research was developed in the scope of the development of a SC simulation model. However, it is the authors' strong conviction that the approach proposed in this paper can also be useful for other domain not necessarily related to SCs, as long as a simulation model of a complex system is being developed in a Big Data context. In fact, SCs are known to be considerably complex systems, which also generate huge amounts of data, thus, the insights obtained in this research were potentialized. Therefore, it is expected that simulation may also help in validating the semantics of data of other systems, which can be useful even in projects aiming to use alternative methods, e.g., feed data to machine learning algorithms.

A particular case of the data issues identified in this research is concerned with the absence of relevant data sources, which, as discussed in subsection 4.1 can have several causes. In this regard, simulation may play two relevant

roles. In a first instance, it can help in understanding the need for a given business process to be included in the model and hence the associated data source. Thereafter and due to the absence of data that is detected, it can then be used to estimate such data, e.g., historical stock levels and production times and capacity.

Finally, this research also showed that, despite current trends emphasizing the need to couple simulation and Big Data technologies, especially for SC problems [4]–[6], considerable problems can still be identified in the data, some of which are hard to identify. This was observed, even though the research was conducted in an excellence environment, where market leader technologies (e.g., SAP) are used and reference business processes in the industry sector are adopted. Thus, the authors argue that considerable efforts must still be made towards improving the quality of data. In a sense, this is expected to be achieved when the recent trends related to Industry 4.0 are fulfilled, e.g., the ability to automatically have interconnected things generating data to be stored and integrated in the BDW for later processing.

6 CONCLUSIONS

SCs are known to be complex systems where simulation may play a relevant role in analyzing such networks and in improving the SCM. Furthermore, such networks are known to comprise several data sources that generate data at increasingly higher volumes, velocities and variety, in what is known as a Big Data environment. Hence, it is expected that by coupling both simulation and Big Data concepts and technologies, the analysis and decision-making process in SCs is further leveraged. Having developed such solution using real data form an automotive electronics SC, it was found that simulation allowed to identify several data issues that were not determined with traditional data profiling techniques. Thus, this paper proposed using simulation as a semantic data validator. This paper also proposed a classification for the types of issues that were faced in a Big Data context and quantified their impact in the volume of data without semantic issues that is obtained. Some of the experiences, lessons learned, and conclusions withdrawn from working in such project were provided in this paper.

Previous studies have addressed the multiple types of problems that exist in the real data. However, the novelty of this research is concerned with the use of simulation to improve typically used data profiling techniques, as well as the fact that it was conducted in a Big Data context, characterized by having multiple data sources (sometimes for the same business process) and having applied Big Data concepts and technologies. In fact, the discovery of some of the issues mentioned in this paper is hindered by such Big Data environment, hence, simulation helped in the identification of those problems. Therefore, it is expected that future researches in the same domain will experience similar issues, with the authors hoping that they find the contributions, insights and experiences shared in this paper to be useful.

The benefits of using simulation as a semantic validator of data are tied with its ability to transform real data into dynamic representations of the real system, where entities represent the flows of materials and information as is described by the data. Thus, certain patterns may be detected, either by visualizing the model animation or by analyzing obtained results. Furthermore, the systems thinking that is required in these cases, as well as the need to integrate and model the data according to simulation needs also heightens the ability to detect errors that traditional data profiling approaches would struggle to identify. For instance, in this research, it was found that the data related

to warehouse movements did not follow the storage strategy of the plant. Conversely, when using alternative approaches, such as feeding data to machine algorithms, data is fed to them and wrong generalizations may be established. Therefore, even such methods could potentially benefit from the approach provided in this paper, as it could validate the data model they are using. It is, indeed, the authors' strong conviction that future SC simulation projects in Big Data contexts will require similar approaches to the one provided in this paper.

This paper also showed that, despite current trends emphasizing the need to couple simulation and Big Data, since the volume of data is huge, there is also a considerable volume that cannot be used due to the several problems it has. In addition, and despite such high volumes of data, other problems like the lack of access to sensitive data also deprive simulation from data that is required. This results in the need to combine the real data provided by Big Data structures with the traditional simulation approach of using statistical distributions to model certain processes.

In terms of future work, the following directions are highlighted. In what concerns the issue of missing historical data, the BDW can be used to maintain such historical values, however, these will only be accessible in the mid- to long-term. The remaining missing data sources have to be covered with solutions aligned with the organization. Lastly, despite the strong conviction that the approach of using simulation as a semantic validator of data, in Big Data contexts, can be applied in other domains, it is still necessary to verify such possibility with future research.

REFERENCES

[1] D. Simchi-Levi, P. Kaminsky, E. Simchi-Levi, and R. Shankar, *Designing and managing the supply chain: concepts, strategies and case studies*. Tata McGraw-Hill Education, 2008.

[2] M. Jahangirian, T. Eldabi, A. Naseer, L. K. Stergioulas, and T. Young, "Simulation in manufacturing and business: A review," *Eur. J. Oper. Res.*, vol. 203, no. 1, pp. 1–13, 2010.

[3] B. Pires *et al.*, "A Bayesian Simulation Approach for Supply Chain Synchronization," in *Proceedings of the 2016 Winter Simulation Conference*, 2016, pp. 3698–3699.

[4] R. Y. Zhong, S. T. Newman, G. Q. Huang, and S. Lan, "Big Data for supply chain management in the service and manufacturing sectors: Challenges, opportunities, and future perspectives," *Comput. Ind. Eng.*, vol. 101, pp. 572–591, 2016.

[5] A. A. Vieira, L. M. Dias, M. Y. Santos, G. A. Pereira, and J. A. Oliveira, "Setting an industry 4.0 research and development agenda for simulation – A literature review," *Int. J. Simul. Model.*, vol. 17, no. 3, pp. 377–390, 2018.

[6] S. Tiwari, H. M. Wee, and Y. Daryanto, "Big data analytics in supply chain management between 2010 and 2016: Insights to industries," *Comput. Ind. Eng.*, vol. 115, pp. 319–330, Jan. 2018.

[7] H. Kagermann, J. Helbig, A. Hellinger, and W. Wahlster, *Recommendations for Implementing the Strategic Initiative INDUSTRIE 4.0: Securing the Future of German Manufacturing Industry ; Final Report of the Industrie 4.0 Working Group*. Forschungsunion, 2013.

[8] H. Lasi, P. Fettke, H.-G. Kemper, T. Feld, and M. Hoffmann, "Industry 4.0," *Bus. Inf. Syst. Eng.*, vol. 6, no. 4, pp. 239–242, 2014.

[9] C. Costa and M. Y. Santos, "Evaluating Several Design Patterns and Trends in Big Data Warehousing Systems," in *30th International Conference on Advanced Information Systems Engineering, CAiSE 2018, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10816 LNCS, 2018, pp. 459–473.

[10] L. M. S. Dias, A. A. C. Vieira, G. A. B. Pereira, and J. A. Oliveira, "Discrete simulation software ranking — A top list of the worldwide most popular and used tools," in *2016 Winter Simulation Conference (WSC)*, 2016, pp. 1060–1071.

[11] A. Vieira, L. M. S. Dias, G. Pereira, and J. A. Oliveira, "Comparison of SIMIO and ARENA simulation tools," in *12th Annual Industrial Simulation Conference (ISC2014)*, 2014, pp. 5–13.

[12] A. A. C. Vieira, L. M. S. Dias, G. A. B. Pereira, J. A. Oliveira, M. Do Sameiro Carvalho, and P. Martins, "Simulation model generation for warehouse management: Case study to test different storage strategies," *Int.*

J. Simul. Process Model., vol. 13, no. 4, pp. 324–336, 2018.
[13] E. Bottani, "Reengineering, simulation and data analysis of an RFID system," *J. Theor. Appl. Electron. Commer. Res.*, vol. 3, no. 1, pp. 13–29, 2008.
[14] C. Gupta, A. Mehta, S. Wang, and U. Dayal, "Fair, effective, efficient and differentiated scheduling in an enterprise data warehouse," in *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology, EDBT'09*, 2009, pp. 696–707.
[15] J. F. Ehmke, D. Großhans, D. C. Mattfeld, and L. D. Smith, "Interactive analysis of discrete-event logistics systems with support of a data warehouse," *Comput. Ind.*, vol. 62, no. 6, pp. 578–586, 2011.
[16] Y. Li and K. D. Joshi, "Data Cleansing Decisions: Insights from Discrete-Event Simulations of Firm Resources and Data Quality," *J. Organ. Comput. Electron. Commer.*, vol. 22, no. 4, pp. 361–393, 2012.
[17] S. S. Nageshwaraniyer, C. Meng, A. Maghsoudi, Y.-J. Son, and S. Dessureault, "Simulation-based decision support system for sustainable coalmining operations," in *62nd IIE Annual Conference and Expo 2012*, 2012, pp. 1574–1583.
[18] E. Kugu, L. Altay, and O. K. Sahingoz, "Using Agent Based Modeling and Simulation for Data Mining," in *19th International Conference on Neural Information Processing, ICONIP, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7664 LNCS, no. PART 2, 2012, pp. 258–265.
[19] T. M. Truong *et al.*, "An implementation of framework of business intelligence for agent-based simulation," in *Proceedings of the Fourth Symposium on Information and Communication Technology - SoICT '13*, 2013, pp. 35–44.
[20] T. M. Truong, F. Amblard, B. Gaudou, and C. S. Blanc, "To calibrate & validate an agent-based simulation model : An application of the combination Framework of BI solution & multi-agent platform," in *ICAART 2014 - Proceedings of the 6th International Conference on Agents and Artificial Intelligence*, 2014, vol. 2, pp. 172–183.
[21] T. M. Truong, F. Amblard, B. Gaudou, and C. S. Blanc, "CFBM - A Framework for Data Driven Approach in Agent-Based Modeling and Simulation," in *2nd International Conference on Nature of Computation and Communication, ICTCC, Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*, vol. 168, 2016, pp. 264–275.
[22] M. Rabe and F. Dross, "A Reinforcement Learning approach for a Decision Support System for logistics networks," in *2015 Winter Simulation Conference (WSC)*, 2015, pp. 2020–2032.
[23] P. Zikopoulos and C. Eaton, *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media, 2011.
[24] S. Madden, "From databases to big data," *IEEE Internet Comput.*, vol. 16, no. 3, pp. 4–6, 2012.
[25] E. Costa, C. Costa, and M. Y. Santos, "Efficient big data modelling and organization for hadoop hive-based data warehouses," in *European, Mediterranean, and Middle Eastern Conference on Information Systems, EMCIS 2017, Lecture Notes in Business Information Processing*, 2017, vol. 299, pp. 3–16.
[26] C. Costa and M. Y. Santos, "The SusCity big data warehousing approach for Smart Cities," in *ACM International Conference Proceeding Series*, 2017, vol. Part F1294, pp. 264–273.
[27] J. R. Lourenço, V. Abramova, M. Vieira, B. Cabral, and J. Bernardino, "NoSQL Databases: A Software Engineering Perspective," 2015, pp. 741–750.
[28] P. Grover and A. K. Kar, "Big Data Analytics: A Review on Theoretical Contributions and Tools Used in Literature," *Glob. J. Flex. Syst. Manag.*, vol. 18, no. 3, pp. 203–229, 2017.
[29] S. Mohanty, M. Jagadeesh, and H. Srivatsa, *Big data imperatives: Enterprise 'Big Data'warehouse,'BI'implementations and analytics*. Apress, 2013.
[30] R. G. Goss and K. Veeramuthu, "Heading towards big data building a better data warehouse for more data, more speed, and more users," in *ASMC 2013 SEMI Advanced Semiconductor Manufacturing Conference*, 2013, pp. 220–225.
[31] A. Thusoo *et al.*, "Hive - a petabyte scale data warehouse using Hadoop," in *2010 IEEE 26th International Conference on Data Engineering (ICDE 2010)*, 2010, pp. 996–1005.
[32] A. Thusoo *et al.*, "Data warehousing and analytics infrastructure at facebook," in *Proceedings of the 2010 international conference on Management of data - SIGMOD '10*, 2010, p. 1013.
[33] E. Costa, C. Costa, and M. Y. Santos, "Evaluating partitioning and bucketing strategies for Hive-based Big Data Warehousing systems," *J. Big Data*, vol. 6, no. 1, p. 34, Dec. 2019.
[34] M. Y. Santos *et al.*, "A Big Data system supporting Bosch Braga Industry 4.0 strategy," *Int. J. Inf. Manage.*, vol. 37, no. 6, pp. 750–760, 2017.

[35] N. Nodarakis, S. Sioutas, A. Tsakalidis, and G. Tzimas, “Using Hadoop for Large Scale Analysis on Twitter: A Technical Report,” *arXiv Prepr. arXiv1602.01248*, Feb. 2016.
[36] R. S. Kv and N. P. Kavya, “Trend analysis of e-commerce data using Hadoop ecosystem,” *Int. J. Comput. Appl.*, vol. 147, no. 6, pp. 1–5, 2016.
[37] J.-H. Thun and D. Hoenig, “An empirical analysis of supply chain risk management in the German automotive industry,” *Int. J. Prod. Econ.*, vol. 131, no. 1, pp. 242–249, 2011.
[38] S. A. Masoud and S. J. Mason, “Integrated cost optimization in a two-stage, automotive supply chain,” *Comput. Oper. Res.*, vol. 67, pp. 1–11, 2016.
[39] A. Ghadge, S. Dani, and R. Kalawsky, “Supply chain risk management: Present and future scope,” *Int. J. Logist. Manag.*, vol. 23, no. 3, pp. 313–339, 2012.
[40] D. Simchi-Levi *et al.*, “Identifying risks and mitigating disruptions in the automotive supply chain,” *Interfaces (Providence).*, vol. 45, no. 5, pp. 375–390, 2015.
[41] A. A. C. Vieira, L. Pedro, M. Y. Santos, J. M. Fernandes, and L. S. Dias, “Data Requirements Elicitation in Big Data Warehousing,” in *European, Mediterranean, and Middle Eastern Conference on Information Systems, EMCIS 2018, Lecture Notes in Business Information Processing*, vol. 341, 2018, pp. 106–113.
[42] G. Popovics, A. Pfeiffer, and L. Monostori, “Generic data structure and validation methodology for simulation of manufacturing systems,” *Int. J. Comput. Integr. Manuf.*, vol. 29, no. 12, pp. 1272–1286, 2016.
[43] R. G. Sargent, “Validation and verification of simulation models,” in *Winter Simulation Conference Proceedings*, 1999, vol. 1, pp. 39–48.
[44] J. Correia, M. Y. Santos, C. Costa, and C. Andrade, “Fast Online Analytical Processing for Big Data Warehousing,” in *2018 International Conference on Intelligent Systems (IS)*, 2018, pp. 435–442.
[45] P. Domingos, “A few useful things to know about machine learning,” *Commun. ACM*, vol. 55, no. 10, p. 78, Oct. 2012.
[46] M. Schmidt, W. Hartmann, and P. Nyhuis, “Simulation based comparison of safety-stock calculation methods,” *CIRP Ann. - Manuf. Technol.*, vol. 61, no. 1, pp. 403–406, 2012.
[47] R. Y. Wang and D. M. Strong, “Beyond Accuracy: What Data Quality Means to Data Consumers,” *J. Manag. Inf. Syst.*, vol. 12, no. 4, pp. 5–33, Mar. 1996.
[48] N. Laranjeiro, S. N. Soydemir, and J. Bernardino, “A Survey on Data Quality: Classifying Poor Data,” in *2015 IEEE 21st Pacific Rim International Symposium on Dependable Computing (PRDC)*, 2015, pp. 179–188.
[49] J. Bokrantz, A. Skoogh, D. Lämkull, A. Hanna, and T. Perera, “Data quality problems in discrete event simulation of manufacturing operations,” *Simulation*, vol. 94, no. 11, pp. 1009–1025, Nov. 2018.
[50] S. Kaisler, F. Armour, J. A. Espinosa, and W. Money, “Big Data: Issues and Challenges Moving Forward,” in *2013 46th Hawaii International Conference on System Sciences*, 2013, pp. 995–1004.