Simulation Modelling Practice and Theory 97 (2019) 101956

Contents lists available at ScienceDirect

Simulation Modelling Practice and Theory

journal homepage: www.elsevier.com/locate/simpat

Supply chain hybrid simulation: From Big Data to distributions and approaches comparison

Antonio A. C. Vieira*, Luís M. S. Dias, Maribel Y. Santos, Guilherme A. B. Pereira, José A. Oliveira

ALGORITMI Research centre, University of Minho 4710-057, Gualtar, Braga, Portugal

ARTICLE INFO

Keywords:
Supply chain
Simulation
Big Data
Industry 4.0

ABSTRACT

The uncertainty and variability of Supply Chains paves the way for simulation to be employed to mitigate such risks. Due to the amounts of data generated by the systems used to manage relevant Supply Chain processes, it is widely recognized that Big Data technologies may bring benefits to Supply Chain simulation models. Nevertheless, a simulation model should also consider statistical distributions, which allow it to be used for purposes such as testing risk scenarios or for prediction. However, when Supply Chains are complex and of huge-scale, performing distribution fitting may not be feasible, which often results in users focusing on subsets of problems or selecting samples of elements, such as suppliers or materials. This paper proposed a hybrid simulation model that runs using data stored in a Big Data Warehouse, statistical distributions or a combination of both approaches. The results show that the former approach brings benefits to the simulations and is essential when setting the model to run based on statistical distributions. Furthermore, this paper also compared these approaches, emphasizing the pros and cons of each, as well as their differences in computational requirements, hence establishing a milestone for future researches in this domain.

1. Introduction

Supply Chains (SCs) are comprised of entities, such as suppliers and customers, wherein material and information exchanges occur, driven by demand and supply interactions [1]. Activities such as production and transportation of raw materials occur in such networks, with the ultimate goal of each entity fulfilling their customers' orders at a minimum cost, whilst improving their competitiveness. In other words, to efficiently manage raw materials receipt and timely schedule deliveries at the right time, place and quantities.

SC systems generate data at increasingly higher rates, volumes and formats, in what is known as the three main characteristics of Big Data [2], In fact, according to Madden [3], this environment in which data is too big, too fast and too hard for existing tools to process, paved the way for the advent of alternative structures to store and integrate data in Big Data contexts. In light of this, Costa and Santos [4] proposed a Big Data Warehouse (BDW) structure, which is 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, machine learning, deep learning and simulations, hence allowing Big Data Analytics (BDA) to be employed.

* Corresponding author.
E-mail addresses: antonio.vieira@dps.uminho.pt (A.A. . Vieira), lsd@dps.uminho.pt (L.M. . Dias), maribel@dsi.uminho.pt (M.Y. Santos), gui@dps.uminho.pt (G.A. . Pereira), zan@dps.uminho.pt (J.A. Oliveira).

https://doi.org/10.1016/j.simpat.2019.101956
Received 16 April 2019; Received in revised form 25 July 2019; Accepted 31 July 2019
Available online 31 July 2019
1569-190X/ © 2019 Elsevier B.V. All rights reserved.

In what concerns simulation, as Jahangirian et al. [5] and Pires et al. [6] postulated, it has been widely used to model SC processes by feeding these models with statistical distributions, however, the use of transactional data to feed such models has not been extensively explored. A fortiori and to the best of the authors' knowledge, no study has used Big Data technologies to model SC systems, despite the benefits that Big Data technologies are expected to bring to simulation of SCs, in accordance with Industry 4.0 and as discussed by several studies [7–11].

Simulation can be used for several purposes, e.g.: visualize the dynamics of systems, determine solutions, understanding complex problems, testing alternative scenarios (e.g., test the impact of certain disruption scenarios) and future prediction. Indeed, the use of Big Data and Simulation, when applied to SC systems, can bring considerable benefits, as it would allow to consider real data originated from multiple data sources and evaluate the response of the system to uncertainty and variable scenarios. However, in the case of using simulation to test alternative scenarios or to make predictions, the use of simulation to reproduce the movements of materials and information that are stored in Big Data structures is not enough, as it would be necessary to complement this with statistical distributions that allow the model to reproduce behavior that has not happened, and thus there is no data with such information.

For complex SCs of huge-scale, to obtain a simulation model that runs based on statistical distributions and is able to mimic the behavior of the system is a complex task. The grand challenge for this task consists in establishing proper statistical distributions for thousands of materials ordered to hundreds of suppliers or by different customers, with adequate parameters for each one. For instance, a given product may be ordered twice a month, while another one may be ordered twice a year; one may come from Asia by sea with a high lead time and another may come through aircraft with a shorter lead time; some materials are single-sourced while others are not. There are numerous possible combinations, resulting in multiple distributions required for the system. Nevertheless, this reality should still be reflected in the simulation model, otherwise a reliable mimic of the system is not achieved. Despite this, often, users simplify the problem by selecting subsets of problems or samples of elements such as materials or suppliers, thus, disregarding the complete view of the system that would allow the complete virtualization of the SC [10,11].

In light of the above discussed, the objectives of this paper are threefold. First, given the absence of studies that have provided modeling approaches for SC simulation models in Big Data contexts, this paper proposes one, which consisted in developing a BDW to store, integrate and provide real industrial data, from an automotive electronics SC, to a simulation model. Second, the approach that was followed to allow the same simulation model to also run based on statistical distributions is provided, which culminated in a brief display of the results that can be obtained when using the tool for future scenario prediction. Hence, the model is able to run with either approach, or even with a combination of both approaches, i.e., it is a hybrid simulation model. Third, the paper presents a short comparison of both approaches in terms of the computational resources required by both, hence serving as a milestone for future works of SC simulations in Big Data contexts.

This paper is structured as follows. Section 2 analyzes works related with the use of simulation in SCs (differentiating between those that used statistical distributions and those that used transactional data in their models) and the existing structures to work in Big Data contexts, while also highlighting existing gaps. Third section presents the framework defined for this project. Afterwards, fourth section starts by presents a brief description of the SC at hand, to better convey the complexity associated with the problem. Thereafter, the section discusses the main development approaches that allow the simulation model to run based on the data stored in the BDW, on statistical distributions or as a combination of both approaches, ending the section with a comparison of these approaches. Finally, conclusions and future work are discussed in the last section.

2. Related work

This section comprises two arts. The first highlights the emphasis that current state of the art is putting on the need for Big Data technologies to be applied in managerial studies in SCs. Next, second subsection presents and discusses the available alternatives for data warehousing, specially focusing on those appropriate for Big Data contexts.

2.1. Simulation in supply chains

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. [10]. 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 crisis scenarios in SCs. In fact, simulation has been extensively applied in SC problems, however, most studies use statistical distributions to model the operations occurring in these networks [5,6]. For examples of such studies, see the studies of Cha-Ume and Chiadamrong [12], Longo and Mirabelli [13], Lee et al. [14], Chen et al. [15], Finke et al. [16], Schmitt and Singh [17], Blanco et al. [18] and Mishra and Chan [19].

Conversely and according to Jahangirian et al. [5], the absence of using real industrial data in simulation models may result in reduced stakeholders interest, with the cited authors noting that there is a gap of simulation studies making use of transactional data in simulation models. For examples of studies using transactional data in simulation models, consider the case of Cheng et al. [20] who used GBSE (General Business Simulation Environment) to help making tactical level decisions in a SC. Schwede et al. [21] developed a SC simulation model of the automotive industry using OTD-NET. The purpose was to use the tool to help in entering in emergent markets. In their turn, Fornasiero et al. [22] and Macchion et al. [23] proposed a simulation model in SIMIO [24] which assessed the impact of orders size in the SC performance. Fornasiero et al. [22] applied their simulation model to a the fashion industry comprised of 60 manufacturers and 1 manufacturer, whilst Macchion et al. [23] applied it to a SC of the footwear industry

comprised of 4 suppliers, 1 warehouse, 1 manufacturer, 1 distributor and 2 customers. In both studies, the authors reported their simulation models are able to retrieve data from the Enterprise Resource Planning (ERP) system. Sahoo and Mani [25] presented a simulation model in ExtendSim to model a SC of the biomass industry. The modelled SC comprised producer and farmer of biomass and suppliers which transported the raw materials to the plant that could store them or process them for later bioenergy production, in order to deliver heat and electricity to customers. The simulation model, among other operational data, stored weather data for long time periods. Ponte et al. [26] evaluated the impact that inventory management and different forecast methods have on the demand variation propagation upstream the SC. The obtained results provided evidence that the efficiency of each inventory model in reducing the bullwhip effect upstream the SC depends on its position on the network.

A fortiori and to the best of the authors' knowledge, no study has used Big Data technologies to model SC systems, which is corroborated by several studies [7–9]. Despite the lack of such studies, its importance has been widely recognized [7–11]. Aligned with the Industry 4.0 philosophy, Kagermann et al. [10] noted the importance of using Big Data in conjunction with Big Data structures, as it allows data from several data sources to be considered in the model, with the associated benefits.

In its turn, Vieira et al. [8] reviewed simulation studies closely related with the concept of Industry 4.0, in order to identify the boiling research directions for simulation, which are aligned with the industrial revolutionary movement. According to the authors, such studies include the use of Big Data 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 offers.

Zhong et al. [7] outlined the current movements on the application of Big Data for Supply Chain Management. According to the authors, the increasing volume of data in the several SC sectors is a challenge which 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.

Lastly, according to Tiwari et al. [9], 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 use in conjunction with such analytics methods (e.g. simulation). In particular, the authors stress the importance of such duo 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.

2.2. Big Data warehousing

Operational databases (DBs) were the mainstream until the mid-1980s. These DBs store operational data, which is involved in daily management processes. In this type of DB, data models are usually drawn using normalization mechanisms so that the physical DB is optimized for inserting, updating or deleting records, ensuring the consistency of the data while saving storage space [27].

Meanwhile, organizations started requiring fast and comprehensive access to information for enhanced decision-making. As a consequence of this, Data Warehouses (DWs) [27], started appearing, which, according to Golfarelli and Rizzi [27], are a collection of methods, techniques and tools that support data analysis and help in decision-making. The main difference between a traditional operational DB and a DW is that the latter is oriented towards analytics, with a subject-oriented and non-volatile repository. Plus, the focus is on improving the performance of the system in terms of response time, thus data is often denormalized to improve the performance, which results in additional storage space being required [28].

Notwithstanding, DWs are no longer capable of dealing with today's world of Big Data contexts. Organizations are generating data at increasingly higher rates, volumes and formats, in what is known as the three main characteristics of Big Data [2]. These constraints paved the way for a new type of structure capable of dealing with this new Big Data context, to continue providing organizations with analytical capabilities [29,30].

In an attempt to propose other approaches for Big Data contexts, different types of solutions have been proposed and implemented. As Costa et al. [29] and Costa and Santos [30] suggest, some considered implementing DWs in NoSQL DBs, albeit these solutions are only scaling operational systems (see [31] for a comparison of NoSQL engines). Eventually, SQL on-Hadoop emerged as a more efficient solution for Big Data environments [29,30,32,33]. See [34] for a comparison of Hadoop and other alternative solutions for Big Data contexts and Grover and Kar [32] for a summary of existing Big Data tools, including the Hadoop ecosystem.

Santos et al. [35] 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, like customer quality claims, making possible the analysis of several Key Performance Indicators (KPIs) in this area and was implemented in the Hadoop ecosystem. Nodarakis et al. [36] 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 [37] also used Hadoop to analyze trends of e-commerce web traffic logs.

Hadoop is an ecosystem based on the MapReduce programming model and the Hadoop Distributed File System (HDFS). Several systems are included in it, such as Hive, Impala and others, which are used for different tasks required under Big Data contexts. Hive is widely adopted by many organizations and was created by Facebook as a way to improve the Hadoop query capabilities that were very limiting and not very productive [38]. At Facebook, it is extensively used for reporting, ad hoc querying and analysis [39]. Hive organizes the data in tables (each table corresponding to an HDFS directory), partitions (sub-directories of the table directory) and buckets (segments of files in HDFS). In addition, it has its own query language: the HiveQL (Hive Query Language). 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) [4].

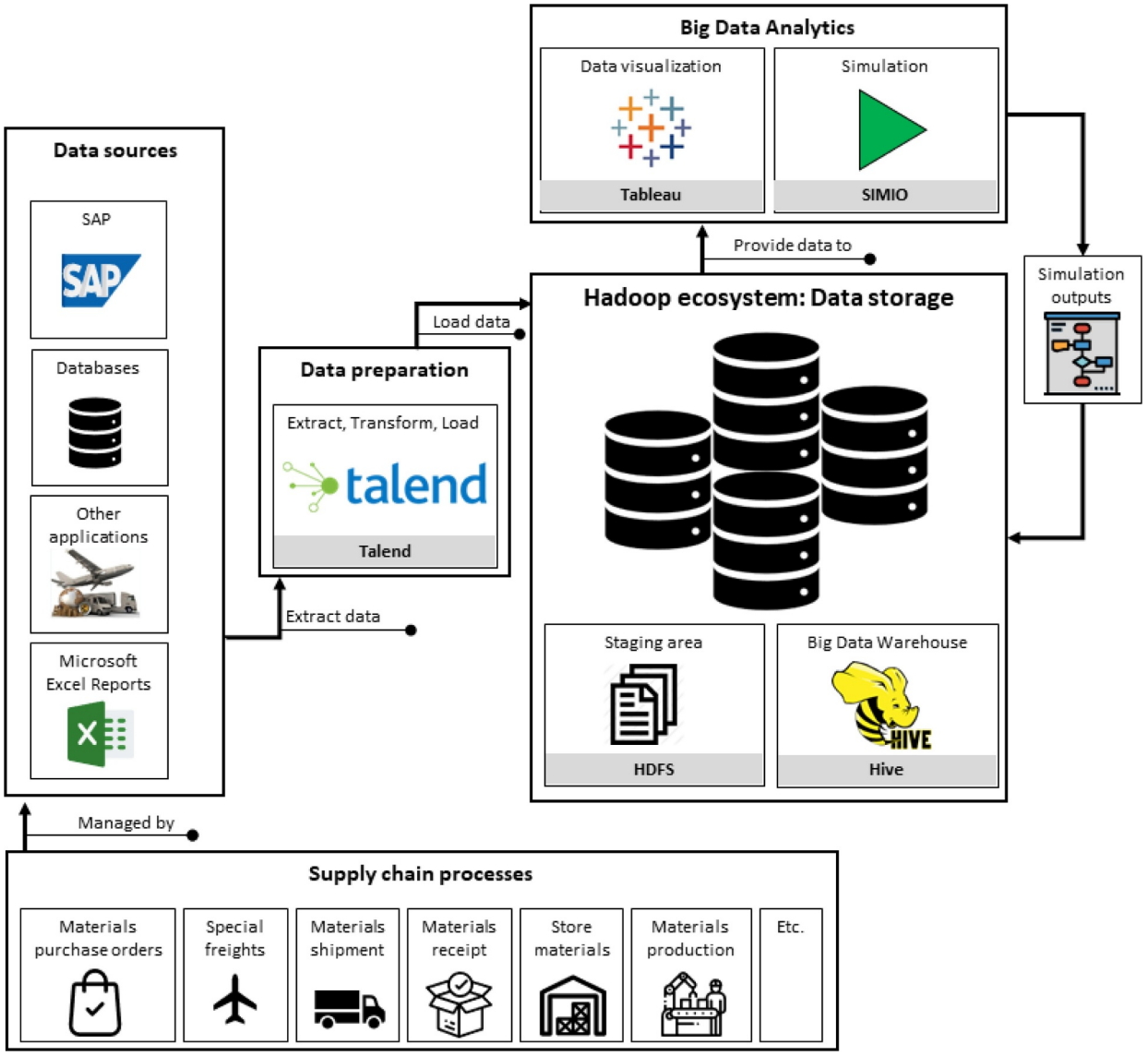

Fig. 1. Proposed framework to develop the BDW and the simulation model.

3. Proposed framework

This section presents the framework defined for this project. In this regard, Fig. 1 depicts such framework. For this development, the Big Data cluster of the Bosch organization was used.

As already indicated, BDWs differ from DWs in many aspects, having, for instance, more flexible schemas than traditional relational ones [40]. However, as Costa and Santos [41] postulated, some similarities between developing traditional DWs and BDWs can be found, namely the business requirements elicitation, the data modeling, the physical design of the BDW tables and the ETL (Extract, Transform, Load) process, or equivalents, e.g., ELT (Extract, Load, Transform).

As illustrated in the lower level of the above figure, the first step considered the identification of the relevant SC processes and their analysis. Raw materials purchase orders, material receipts, production consumptions and master data were some of the main SC processes that were considered. The selection of these processes was the result of several interviews, focus groups and other group sessions with process experts.

After selecting the business processes, the data sources were also selected. It should be noted that, in big organizations the same business process may be managed by different systems. Therefore, to select the most relevant ones, other set of interviews focus groups and group sessions were conducted with process experts, culminating in the inclusion of several tailored software, information systems, ERP, DBs and others. The ERP used at Bosch is SAP, thus, most of the data related with these processes is stored in this ERP. Other data sources included, for instance, tailored software to manage special freights, Microsoft Excel file to manage all the material

receipts at the plant and Microsoft Access DB to manage the early arrivals of materials to the plant. In addition, the data originated from such systems comes in different formats (e.g., Access DB, Excel).

To collect data from these sources, the traditional ETL approach was followed, meaning that data is collected, the necessary transformations are performed and, thereafter, data is sent to the staging area, namely to the HDFS system; Talend was used for this task. The same software was used to create the BDW tables (in Hive), in the Hadoop ecosystem of the Big Data cluster. These tables comprise denormalized data and were modeled according to the simulation needs, so that there is no need to search for any values among different tables, during the simulation.

With these BDW tables, different purposes can be achieved, such as: ad hoc querying, data analytics and visualization (e.g., Tableau) and simulations. In this paper, the SIMIO simulation tool was used. Moreover, the simulation model itself may also generate files with simulation results. These files may also be subject to ETL jobs, to create additional Hive tables with these results, enabling further data analysis or even their integration in the simulation model, which was the case with this paper, as will be afterwards discussed. Alternatively, the SIMIO software can also be used to directly create and load the results to the Hive tables, since it is connected to the BDW.

4. From big data to statistical distributions: approaches and comparison

This section discusses each implemented approach that allows the simulation model to run based on the data stored in the BDW, using statistical distributions or a combination of both approaches. Thus, first subsection provides a brief characterization of the SC at hand. Thereafter, second subsection discusses the approach adopted to allow the model to run by using the data in the BDW, while the third subsection concerns the approach that allows the simulation to run based on statistical distributions, while it also described how both approaches can be combined. Finally, last subsection compares both approaches.

4.1. Supply chain characterization

This subsection briefly describes the SC at hand, which comprises an automotive electronics manufacturer, of the Bosch Group, and its suppliers from all around the world. In this system, around 7000 different types of materials are actively being supplied by roughly 500 different suppliers, located in more than 30 countries. Moreover, Germany, Netherlands, Switzerland, Spain, China, Taiwan and Malasya are the countries that supply more types of materials. Most of the suppliers are from Europe and Asia, with Germany (209 suppliers) and Netherlands (10 suppliers) having more suppliers and shipments from Europe, and Malasya (16 suppliers), Taiwan (13 suppliers), China (12 suppliers), Hong Kong (11 suppliers) and Singapore (7 suppliers) having more shipments from Asia.

Car manufacturers need to comply with very strict security norms for their products, while still providing high levels of product customization, required by increasingly demanding end customers ([43,42]). At the same time, an ordinary car is comprised of multiple materials supplied by single sources, exposing manufacturers to specific suppliers, thereby posing a disruption danger for the entire SC ([44]). Hence, entities interoperating in these SCs need to comply between them, in order not to jeopardize the entire chain ([45]).

4.2. Big Data approach

SC systems comprise flows of external and internal material and information movements. Thus, the entities flowing in the developed simulation model represent these types of movements, i.e., orders to suppliers - and the respective material arrival - and material movements that occur within the plant (e.g., store materials in the warehouse and send them to production). Fig. 2 shows the objects responsible for creating the entities of this simulation model. The “CreateSupplierOrders” Source object is selected and is the object responsible for creating orders to suppliers, while the remaining 4 Sources are used to create material movements that occur within the plant.

The figure shows SMIO objects on the left side, while on the right side, the properties of the selected object highlighted in green are displayed. From the displayed properties, the following can be emphasized:

- Entity Type: defines the type of entity created. The same type of entity is created, regardless of the type of movement (internal or external), since these are differentiated using appropriate symbols;
- Arrival Mode: sets the mode for creating entities. By default, this is set to use the data of the BDW, thus, the mode was set to Arrive Table;
- ArrivalTime Property: defines the column of the Hive table that sets the date at which entities should be created;
- Entities Per Arrival: sets how many entities are created for each arrival. In this case, a Boolean expression was used, which allows the model to stop creating entities using the data of the BDW and change to created them based on statistical distributions;
- Row Number: specifies the data row which is assigned to each created entity. Since the data and the simulation model were modelled so that each row corresponds to a different entity, the created entities is associated to the row with the same date that which trigged its creation, meaning that it retains the attribute values of this row, as its own attributes;
- Created Entity: sets the process that is executed by each entity, when created.

As the object selected in Fig. 2 (CreateSupplierOrders) is responsible for creating entities related with order to suppliers and the subsequent arrivals, the process executed in the Create Entity property models this behavior. Such process is illustrated in Fig. 2. However, for the remaining objects, as they model other behaviors, different processes are executed, e.g.: to model the shipment of

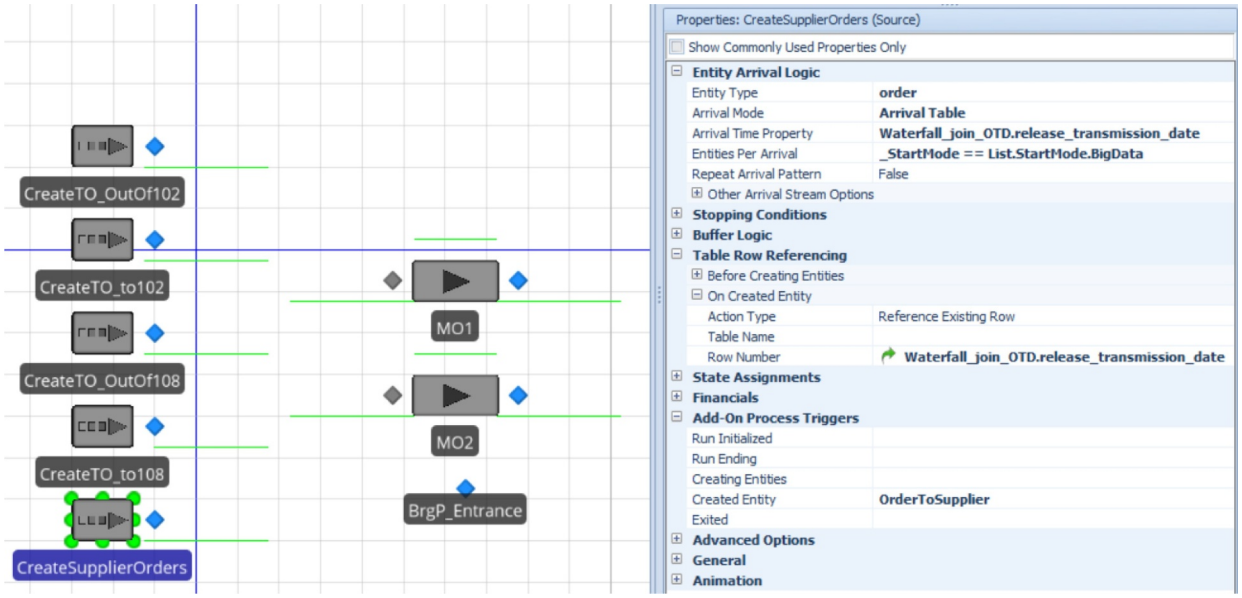

Fig. 2. Properties of the "CreateSupplierOrders" object.

orders from suppliers to the plant and the material movements that occur within the plant see Fig. 3 for an example of a process.

The process depicted in the above figure starts by making some assignments to the created entity, related with its associated data row. Thereafter, the entity is transferred to Free Space, using the "Go to Free Space" Transfer step, and travels to the location of the supplier. This allows entities to travel between objects without requiring any link or connection between them or even move freely in an orthogonal tridimensional space. To achieve this, the movement of entities needs to be modeled using processes, which adds complexity to the development, as connections between two distinct locations are not used. In particular, this movement represents the activity of placing an order to a supplier and was modelled to occur at a very high speed; this allows users to visualize orders being placed. When arriving at the location of the supplier, the entity gets the associated transit and lead time.

When the modeling is complete, it is possible to retrieve results which mimic the material and information movements represented by the data stored in the BDW. In this regard, Fig. 4 shows the total consumed, ordered and arrived quantities and Fig. 5 shows the number of orders placed and received.

4.3. Statistical distribution approach

This subsection addresses the approach that was adopted to allow the model to run based on statistical distributions. In this regard, the following mains steps were followed.

Step 1) Define the processes to model with distributions

The following processes were selected to be modelled as statistical distributions:

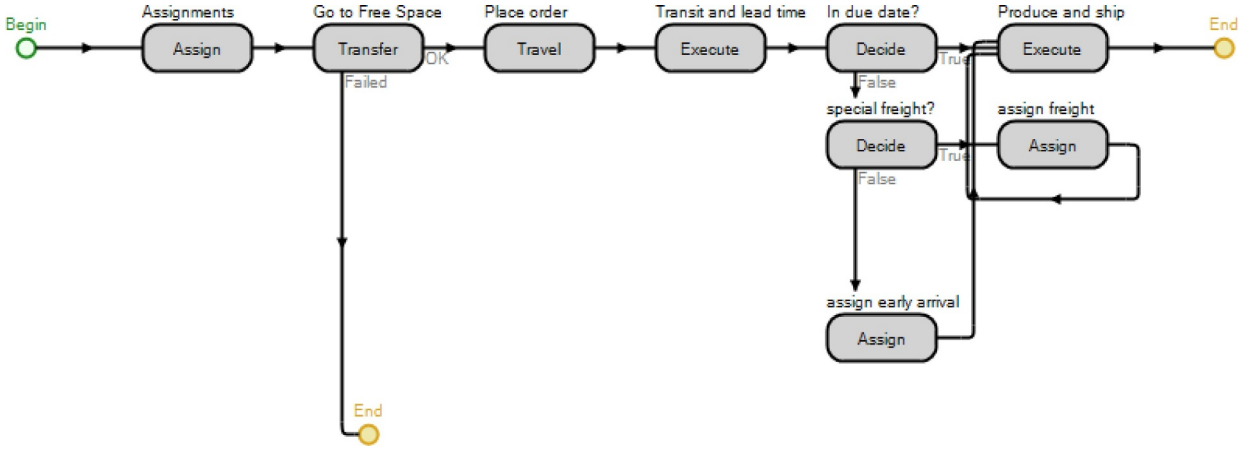

Fig. 3. Process executed to model orders being placed to the respective suppliers.

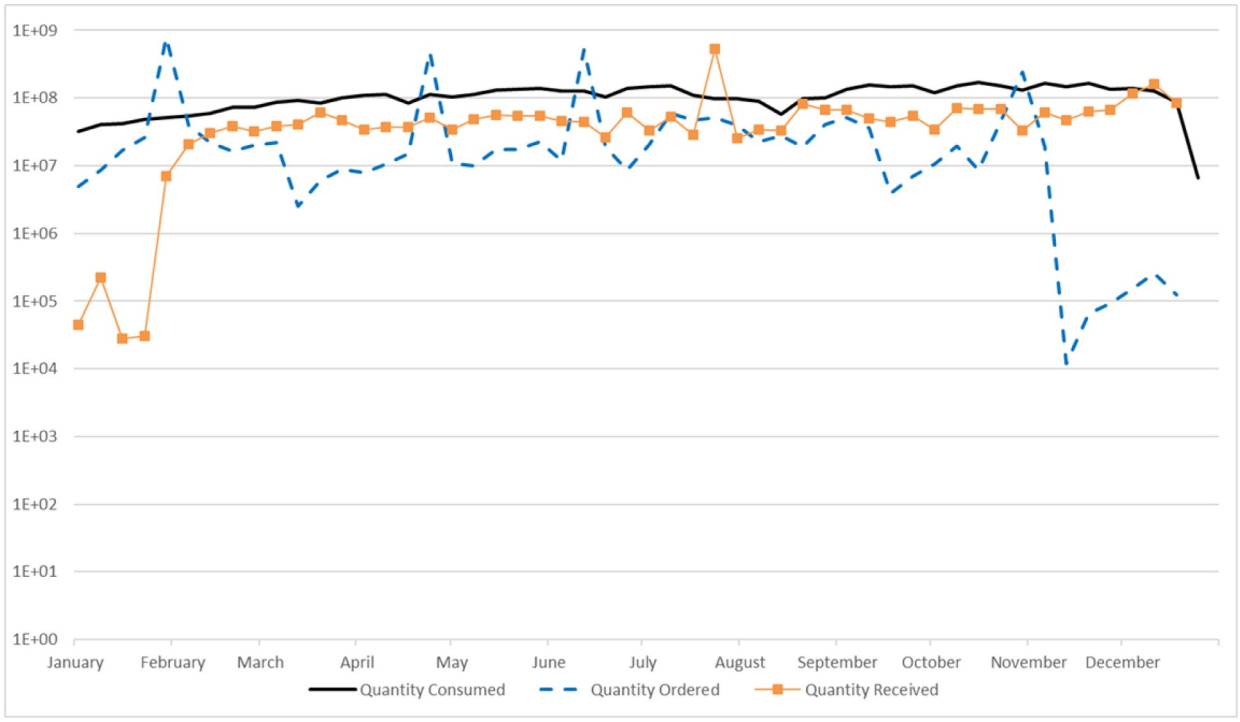

Fig. 4. Total quantity of materials ordered, received and consumed per week.

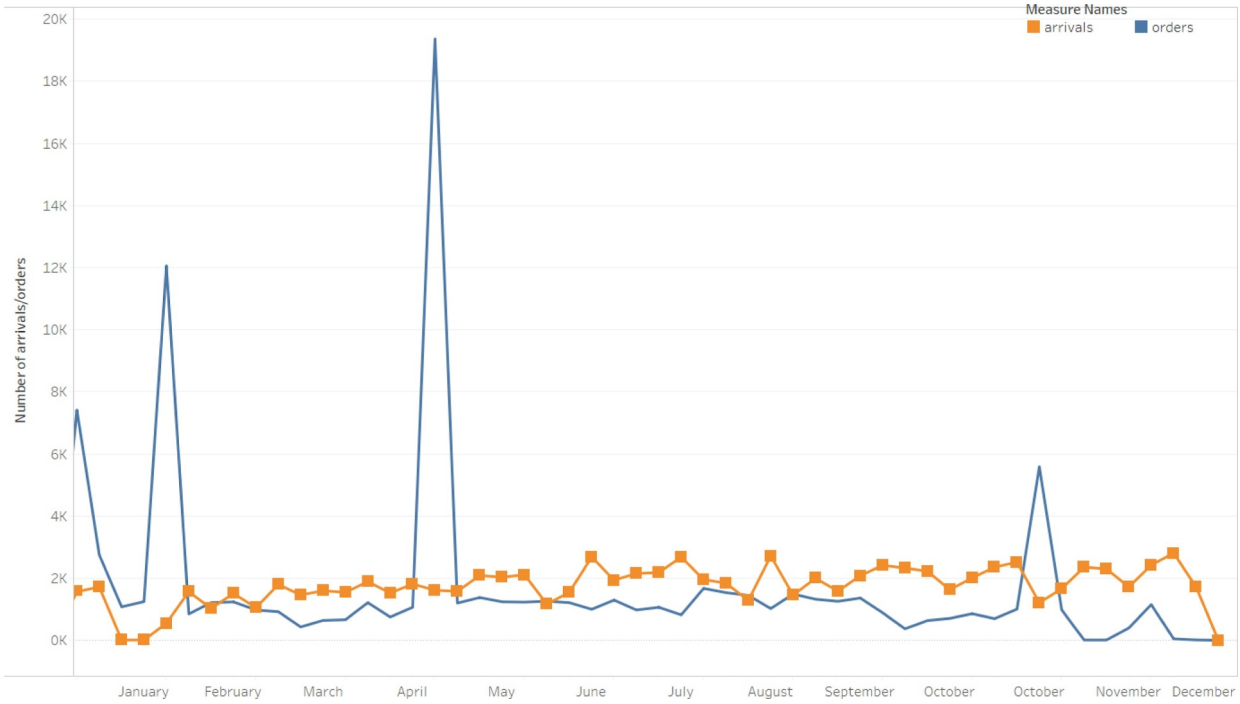

Fig. 5. Number of orders placed and arrivals per week.

- Supplier orders' interarrival time: defines the rate at which orders, for each material, are placed to suppliers;
- Quantity received from supplier;
- Production orders' interarrival time: similar to the supplier orders, this defines the interarrival time between production consumption orders, for each material;

- Quantity ordered to production;
- Lead time per supplier: defines the lead time duration per supplier.

These processes also correspond to the Hive tables that were created in the Big Data approach. In addition, and contrarily to the remaining processes, the lead time cannot be estimated based on other lead times of the same material, since materials may be supplied by different suppliers of different countries or continents. Therefore, the ideal approach would be to estimate the lead time based on other orders of the same material, to the same supplier. However, this would result in samples with few observations, hence compromising the distribution fitting quality. Thus, the lead times were estimated based on other orders, to the same supplier, regardless of the ordered material.

Apart from the above mentioned processes, it is still necessary to use data from the BDW for the static data of elements such as the transit time and supplier for each material, which, in this system is fixed per supplier. Thus, a Hive table was created to store those static values for each material and supplier.

Step 2) Define clusters for materials, suppliers and customers

Defining these clusters is necessary, in order to select samples of materials, suppliers and customers that can, to the possible degree, represent the complete universe of materials and suppliers of the plant. Notwithstanding and contrarily to typical approaches, this one will not maintain these clusters. Rather, these clusters are only selected to determine the distributions more adequate for each process. Afterwards, all the elements will be used in the distribution fitting. For this step, the views of experts from the company was important, as they have the experience to know the most important clusters and those that better represent the SC.

Step 3) Select materials and suppliers based on the defined clusters

For this task, both querying the BDW and interviews with experts was important. In fact, it was relevant to complement both approaches, since they allow to complement data which is not contained in the BDW, e.g., there is no available data for a given process.

Step 4) Using simulation results to create new Hive tables

This step consisted in running the model using the data stored in the BDW, in order to obtain results that are thereafter used to set the distributions parameters, in accordance to what had been depicted in Fig. 1. Thus, for each process described in Step 1) and for each material and supplier selected in Step 3), all registers, as well as other aggregation values, such as the average, standard deviation, minimum, maximum and number of observations are obtained through simulation and stored in the appropriate BDW tables.

In this case, the simulation was used to record the mentioned aggregation values, rather than directly querying the BDW, due to multiple data issues that were faced when working in this project. In fact, such issues required multiple approaches to bypass them. Hence, in these cases, it is important to follow the described approach in this step, which are expected to be common in Big Data contexts.

Step 5) Distribution fitting for the selected processes

This step consists in the distribution fitting phase, which comprises the following steps: First, use the simulation results obtained in the previous step to select the best fitting distribution. The software Arena Input Analyzer was used for this purpose. Fig. 6 shows an example of the distribution fitting conducted to the lead time of a supplier.

A similar graph for each element selected in Step 3) and for each process selected in Step 1) should be done. Second, based on the selected distribution, the distribution parameters must be determined, based on the aggregation values obtained in the previous step, as required by each distribution.

Step 6) Update the simulation model to run based on random distributions

At this point, it was necessary to update to model, so that it could run the model with data from the BDW, using distributions, or a combination of both. In light of this, Fig. 7 shows the process that was used for this.

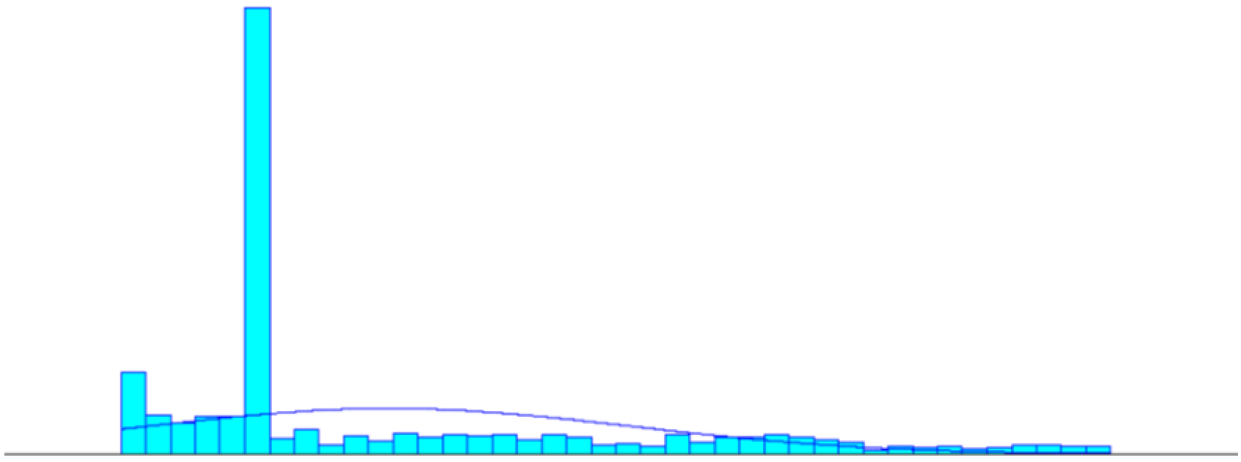

Fig. 6. Histograms of distribution fitting for (bottom) lead time of a supplier selected in Step 3).

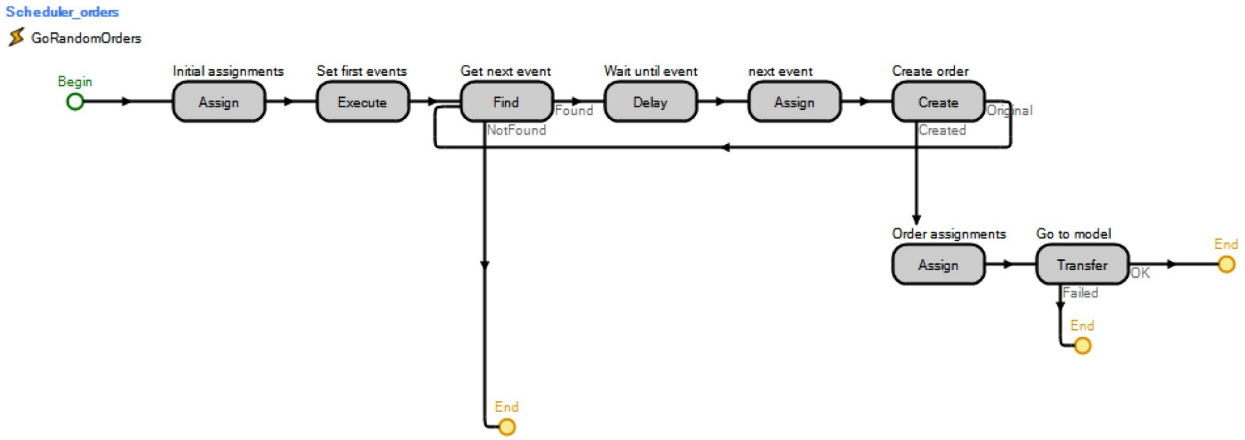

Fig. 7. Process used to create entities based on the determined random distributions.

This process can either be executed at the beginning of the simulation run, if the model is set to run based on statistical distributions, or run from a given simulation time onwards, hence allowing to combine both run modes.

The process starts by turning off the Source objects (see Fig. 2), so that they do not create more entities based on the Hive tables. The process, then, iterates through all materials to calculate the simulation time of their first order and stores it in an array, which has one index for each material. Thereafter, the process selects the next entity to be created, using the Find "Get next event" step, to get the minimum simulation time of all the simulation times previously calculated. Having selected the next order to create, the Delay "Wait until event" step is used to hold the token executing this process until the simulation time retrieved by the Find step, i.e., until the time of the next entity arrival. When the time is reached, the value of the next arrival of the material in question is updated and the Create "Create order" step is used to create an entity referring to the type of material in question. At this point, a new token is created representing the created entity, which exits the Create step through the "created" branch. In its turn, the main token continues the main process, by executing again the Find step, in order to get the next entity to be created. Finally, the new token makes some assignments to the created entity, concerning the order quantity, the type of material and others. The created entity is thereafter sent to the model using the Travel "Go to model" step. With this process, both orders to suppliers and production orders can be created.

Step 7) Run the simulation model based on statistical distributions

Finally, it is possible to run the simulation model using statistical distribution or combine it with the Big Data approach and

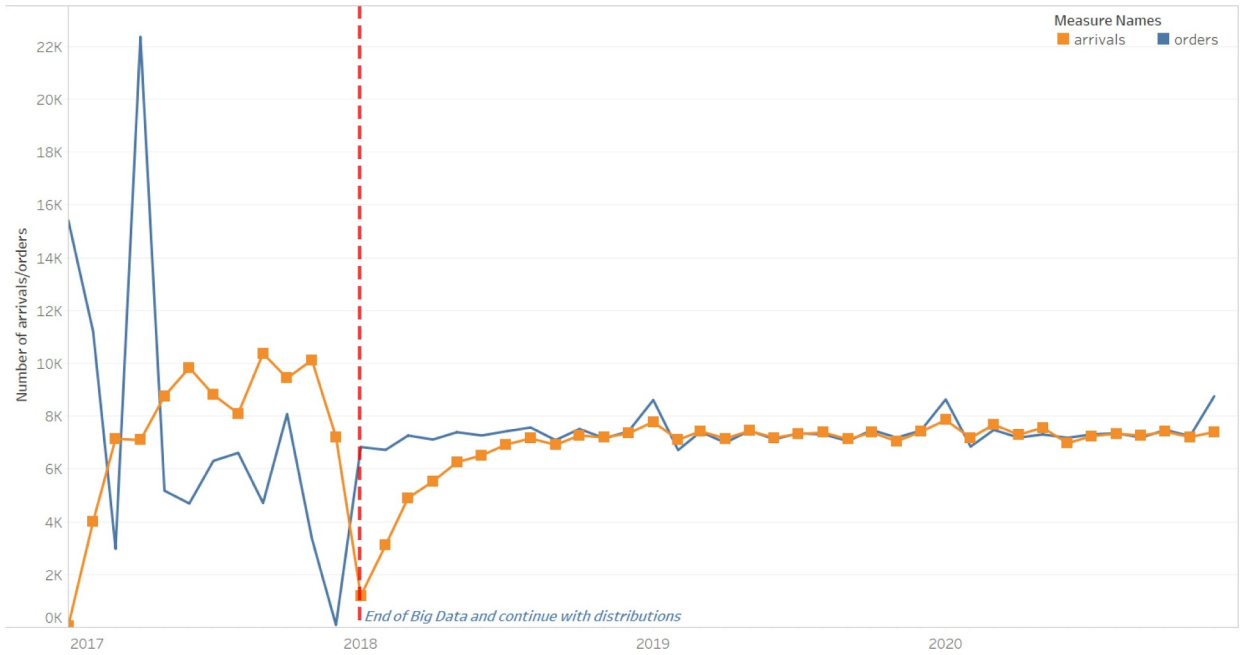

Fig. 8. Number of orders and arrivals per week by using data from the BDW for 2017 and random distributions from 2018 to 2020.

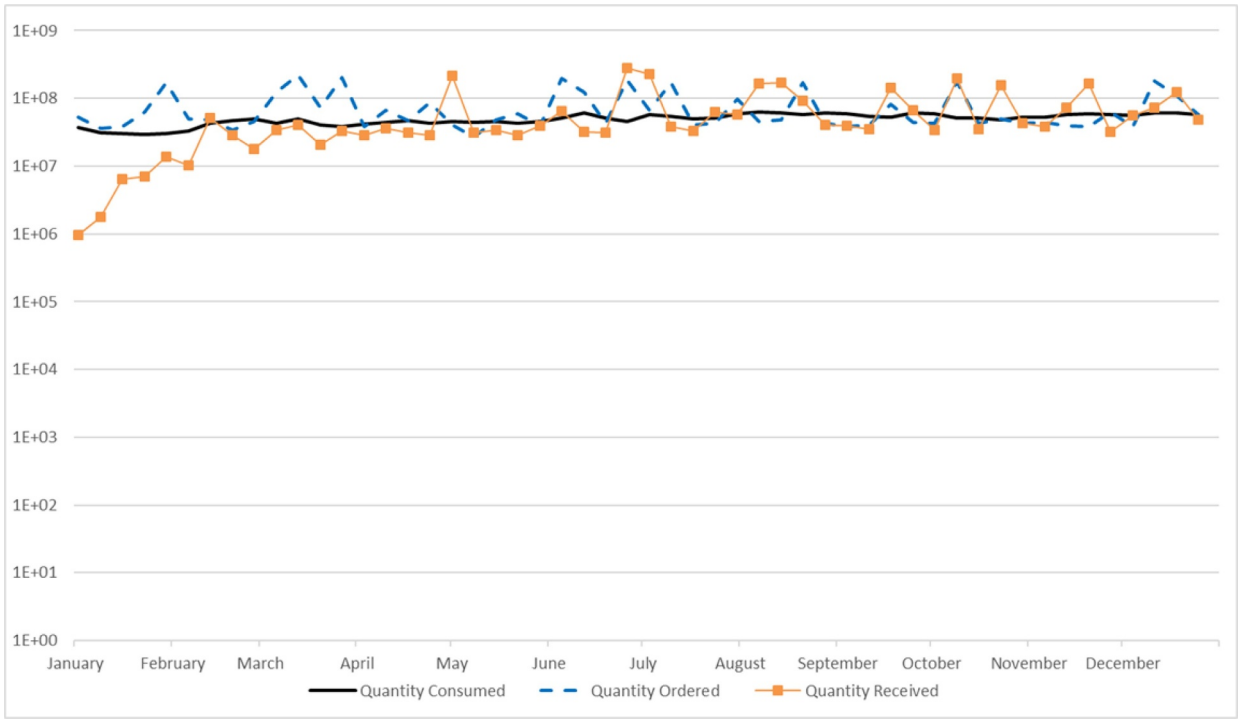

Fig. 9. Total quantities consumed, ordered and arriving at the plant when running the model based on distributions.

analyze the respective results. In this regard, Fig. 8 shows the obtained results for running the model using data from the BDW and statistical distributions for the three following years, and Fig. 9 shows the total consumed, ordered and arrived quantities per week, when running the simulation using statistical distributions.

The approach described in this subsection used the simulation model running based on the data stored in the BDW to determine the adequate statistical distributions, in a semi-automatic way. Alternative approaches would have to either do this manually, which in complex SCs of considerable size may not be feasible, or select samples of SC processes or its elements, e.g. suppliers and materials. Hence, both approaches were necessary, in order to fulfill the objective of running the model with statistical distributions and still capture all the detail of the SC at hand. Next subsection provides a comparison of the modelled approaches.

4.4. Approaches comparison

This subsection compares the simulation approaches of the developed hybrid simulation model. By comparing the results depicted in Figs. 5 and 8, it can be seen that when running the model based on the statistical distributions, the average values remain slightly constant, while the peaks, namely in the number of placed orders, did not occur so frequently, which is the results of the distributions that attenuate the fluctuations. On the other hand, indeed, the approach of using the Big Data provided greater detail to the analysis, as the presented related graphs depicted, in contrast to the ones related with running the simulation with statistical distributions.

Notwithstanding, running simulations based on statistical distributions in a complex and huge-scale SC such as this one, still required running the model based on the data of the BDW, as the approach described in this subsection demonstrated. Apart from the need of the real data stored in the BDW to reach the distributions and their parameters, the approach still required the use of data to maintain the coherence of the model. For instance, when ordering a material, it is necessary to determine available suppliers and their country, in order to place the order to that geographic location; the same applies for material's characteristics, e.g.: standard price, shelf life, safety stock. Furthermore, one of the major benefits of running the simulation based on statistical distributions comes from its ability to allow the model to estimate future scenarios based on data from the past, similar to what was done in Fig. 8. However, this should be coupled with seasonality methods to confer greater realism and coherence to the model. For instance, there are some periods of the year where the number of arrivals decrease, e.g. end of the year.

To compare the approaches in terms of computational resources required by the modelled approaches, the number of created entities, number of instructions executed by the simulation engine, the computer memory required for each experiment and the elapsed time to run, load and save the model were recorded. Table 1 summarizes the results obtained for the modelled approaches: one experiment for each approach.

As can be seen, the benefit of the greater level of detail provided by the Big Data approach (experiment A) resulted in considerable computer memory being required to run a single replication, as 16 GB of memory were used. However, note that this memory was used even though there were data sources which would be relevant for this study, but could be used, for several reasons that not

Table 1
Summary of elapsed time, required memory, executed instructions and created entities per replication of the considered approaches.

Run*	simulation Time	Created Entities	Executed Instructions	Memory Required	Elapsed time(minutes)		
					Running	Loading	Saving
A	1 year	~2 M	~150 M	16 GB	~6	~10	~2
B	1 year	~2 M	~150 M	0.6 GB	~180	< 1	< 1
C	4 years	~8 M	~800 M	16 GB	~600	< 1	< 1

A: Simulation using only the data of the BDW.
B: Simulation using only statistical distributions.
C: Simulation using the data of the BDW for 1 year and statistical distributions for the next 3 years.
M = 1,000,000 entities / GB = GBs.
* Experiments conducted on a desktop computer with 64 bits Windows Server 2016, Intel® Core™ i7-6950X CPU and 64 GB memory, using 64 bits SIMIO simulation software.

discussed in this paper, as it was considered to be out of the scope of this paper. It is expected that, with this data that is missing, the required memory would considerably increase. On the other hand, it is also seen that running the model based on statistical distributions (experiment B), indeed, decreased the necessary memory, albeit the elapsed time to run a single simulation replication considerably increased from 6 min in experiment A to 180 min in experiment B.

5. Conclusions

Being complex and dynamic networks, SCs are prone to uncertain events that may affect its performance. Therefore, proper decision-support systems are required, in order to mitigate the impact of such risks, thus, allowing proactive measures to be taken, rather than reactive ones. For this purpose, simulation may be used, as it allows alternative scenarios to be tested including future scenarios prediction, among other benefits. In light of this, this paper proposed a hybrid simulation model, which allows the simulations to run both based on the data stored in the BDW and on statistical distributions, or in a combination of both approaches.

5.1. General conclusions and managerial implications

With the uprising of Industry 4.0 and the advent of Big Data technologies, it is expected that the huge volumes of data generated at increasingly higher velocities should bring additional benefits and insights, as they provide the realism and detail that the typical simulation approach (using statistical distributions) does not. However, especially in the case of using simulation to test risk scenarios and future prediction, this research has showed that the traditional simulation approach must still be considered. The main reason for this consists in the complexity and size of SCs, which is a typical characteristic. I.e., in a Big Data context where hundreds of agents order, produce and deliver hundreds of materials with different parameters, performing distribution fitting for all these elements would not be feasible. In this regard, the approach provided in this paper uses both the BDW and the simulation model to select the best fitting distribution and their adequate parameters for all the elements of the SC. Hence, the distribution fitting is performed by the simulation model and not by the user. In addition, the data of the BDW must still be considered, in order to maintain the coherence of the simulation model, e.g., to know the suppliers for each material.

Despite this semi-automatic distribution fitting, it is important to note that the view of managers from the company is still of extreme relevance. In fact, while working on this project, and despite being developed in a Big Data context, several data issues were experienced (these were not discussed in this paper as it was out of the scope of the objective established for this paper). One subset of these data issues consists in data sources that could not be used. Thus, to have an efficient view of the SC network, the views of managers are important, as they may provide insights that the stored data does not reveal.

Since this research considered a real case study, the solution presented in this paper, indeed, can only be applied to the plant of the case study. In fact, even plants belonging to a same organization, with strict standardized norms, have their own tailored software and their business processes that do not apply to other plants. Such was the case with this plant. I.e., while most of the data sources used in the project and the considered business processes are standardized among the remaining plants of the organization, there were still some of the mentioned elements that are not standard among the organization. The result is a BDW and a consequent simulation model that are tailored for the plant of the case study.

However, the authors argue that the lack of completely standardized plants (both in terms of data sources and business processes) and the consequent difficulty in developing equally standard SC simulation models in Big Data contexts should not be seen as a handicap. Each organization and each plant have their own socio-economic and geographic factors that must be considered and will inevitably result in differences in the simulation models that aim to produce reliable mimics of these systems. However, if the goal of establishing standard simulation models is a reality, then efforts should be made towards the standardization of data sources and business processes, so that fewer parameters need to be set when setting instances of simulation models for each case.

Notwithstanding the above exposed, this paper hopes to contribute to researchers and practitioners working on similar projects, by sharing the modeling approaches, main faced difficulties and conclusions withdrawn from the experience of working in this project. Furthermore, it is the authors' strong conviction that the approach followed in this paper would also have to be considered by other projects, as long as the system in analysis considers a great number of agents and materials being ordered, produced and

delivered with different parameters. As such cases comprise situations in which manually performing the distribution fitting for all its elements is not feasible, the use of a data repository (in this case, a BDW) and simulation model are necessary, in order to obtain the necessary statistical distributions. Hence, an approach similar to the one provided in this paper would have to be conducted.

5.2. Limitations

Despite the above discussed conclusions, it should be noted that the suggested approach to allow the simulation to run based on statistical distributions resulted in metrics that did not consider the seasonality inherent to the real system. Thus, the system's performance tended to the average values of such processes. However, if seasonality methods are included in the simulation, the metrics based on distributions would alike the metrics of the real system, opening room for more accurate future scenarios projection. Similarly, this research focused on the approach to obtain the parameters for the selected distributions for each selected business process. However, it did not focus on how to select the best fitting distributions for each case. While these aspects can be seen as the major limitations of this study, it should also be noted that these were intentionally left out of the scope of this research, as providing such aspects could deviate the reader from the main objective established for this paper and considerably increase the size of the paper.

5.3. Future research

Finally, despite fulfilling the research objectives for this paper, there is still room for future research, as this research is part of an ongoing project. Such items are now discussed. First, despite using real industrial data, multiple data issues were faced. Some of these issues are related with the organizational data policies. Thus, efforts should be made towards correcting these. Notwithstanding, this leads the authors to conclude that, despite the current trends emphasizing the need to couple SC simulations with Big Data technologies, organizations are still struggling with the quality of their data. This is especially relevant in simulation studies, as these will produce dynamic models based on the data they get. Thus, if relevant data is missing, the simulations may lose reliability. Nevertheless, it is expected that such issues are somewhat mitigated as the Industry 4.0 completely materializes, allowing data to be automatically collected, treated, stored, integrated and provided to simulation models.

Apart from the complexity and size characteristics of SCs that were addressed in this paper, these networks are also known to be quite dynamic, i.e., the relations between agents may frequently change. Such changes may result in the simulation model and the BDW to be no longer accurate. As such, efforts must be made towards allowing both to operate in real-time. Regarding the former, it should automatically adapt to the data stored in BDW, even if data changes occur, using, for instance, data-driven approaches. Regarding the BDW, other Big Data concepts can be applied to allow this real-time feature. In addition, the real-time interoperability between systems exchanging data must also be ensured, e.g. SAP and BDW.

Finally, with the advent of artificial intelligence methods, such as machine learning or deep learning, it seems reasonable to assert that the next window opportunity that simulation practitioners need to capitalize on concerns with the use of these technologies in SC simulation models in Big Data contexts. Hence, using the available Big Data and Artificial Intelligent algorithms to confer actual intelligence to the agents considered in simulations, would allow simulation models to consider not only the individual behavior of agents, but also to observe their actions when they gain “intelligence”.

Acknowledgments

This work has been supported by national funds through FCT – Fundação para a Ciência e Tecnologia within the Project Scope: UID/CEC/00319/2019 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] D. Simchi-Levi, P. Kaminsky, E. Simchi-Levi, R. Shankar, Designing and Managing the Supply chain: concepts, Strategies and Case Studies, Tata McGraw-Hill Education, 2008.
[2] P. Zikopoulos, C. Eaton, Understanding Big data: Analytics for Enterprise Class Hadoop and Streaming Data, McGraw-Hill Osborne Media, 2011.
[3] S. Madden, From databases to big data, IEEE Internet Comput 16 (3) (2012) 4–6.
[4] C. Costa, M.Y. Santos, Evaluating several design patterns and trends in Big Data warehousing systems, 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), 10816 2018, pp. 459–473 LNCS.
[5] M. Jahangirian, T. Eldabi, A. Naseer, L.K. Stergioulas, T. Young, Simulation in manufacturing and business: a review, Eur. J. Oper. Res. 203 (1) (2010) 1–13.
[6] B. Pires, et al., A bayesian simulation approach for supply chain synchronization, Proceedings of the 2016 Winter Simulation Conference, 2016, pp. 3698–3699.
[7] R.Y. Zhong, S.T. Newman, G.Q. Huang, S. Lan, Big data for supply chain management in the service and manufacturing sectors: challenges, opportunities, and future perspectives, Comput. Ind. Eng. 101 (2016) 572–591.
[8] A.A. Vieira, L.M. Dias, M.Y. Santos, G.A. Pereira, J.A. Oliveira, Setting an industry 4.0 research and development agenda for simulation – A literature review, Int. J. Simul. Model. 17 (3) (2018) 377–390.
[9] S. Tiwari, H.M. Wee, Y. Daryanto, Big Data analytics in supply chain management between 2010 and 2016: insights to industries, Comput. Ind. Eng. 115 (2018) 319–330.
[10] H. Kagermann, J. Helbig, A. Hellinger, W. Wahlster, Recommendations for Implementing the Strategic Initiative INDUSTRIE 4.0: Securing the Future of German Manufacturing Industry , Forschungsunion, 2013 Final Report of the Industrie 4.0 Working Group.
[11] H. Lasi, P. Fettke, H.-.G. Kemper, T. Feld, M. Hoffmann, Industry 4.0, Bus. Inf. Syst. Eng. 6 (4) (2014) 239–242.

[12] K. Cha-Ume, N. Chiadamrong, Meta-prediction model for introducing lateral transshipment policies in a retail supply chain network through regression analysis, Eur. J. Ind. Eng. 12 (2) (2018) 199–232.
[13] F. Longo, G. Mirabelli, An advanced supply chain management tool based on modeling and simulation, Comput. Ind. Eng. 54 (3) (2008) 570–588.
[14] Y.M. Lee, S. Ghosh, M. Ettl, Simulating distribution of emergency relief supplies for disaster response operations, Winter Simulation Conference, 2009, pp. 2797–2808.
[15] Y. Chen, L. Mockus, S. Orcun, G.V. R.eklaitis, Simulation-optimization approach to clinical trial supply chain management with demand scenario forecast, Comput. Chem. Eng. 40 (2012) 82–96.
[16] G.R. Finke, A.J. Schmitt, M. Singh, Modeling and simulating supply chain schedule risk, Proceedings of the Winter Simulation Conference, 2010, pp. 3472–3481.
[17] A.J. Schmitt, M. Singh, Quantifying supply chain disruption risk using Monte Carlo and discrete-event simulation, Winter Simulation Conference, 2009, pp. 1237–1248.
[18] E.E. Blanco, X. (Cissy) Yang, E. Gralla, G. Godding, E. Rodriguez, Using discrete-event simulation for evaluating non-linear supply chain phenomena, Proceedings of the Winter Simulation Conference, 2011, pp. 2260–2272.
[19] M. Mishra, F.T.S. Chan, Impact evaluation of supply chain initiatives: a system simulation methodology, Int. J. Prod. Res. 50 (6) (2012) 1554–1567.
[20] F. Cheng, Y.M. Lee, H.W. Ding, W. Wang, S. Stephens, Simulating order fulfillment and supply planning for a vertically aligned industry solution business, Proceedings of the 40th Conference on Winter Simulation, 2008, pp. 2609–2615.
[21] C. Schwede, B. Sieben, Y. Song, B. Hellingrath, A. Wagenitz, A simulation-based method for the design of supply strategies to enter developing markets, Int. J. Simul. Process Model. 5 (4) (2009) 324–336.
[22] R. Fornasiero, L. Macchion, A. Vinelli, Supply chain configuration towards customization: a comparison between small and large series production, IFAC-PapersOnLine 28 (3) (2015) 1428–1433.
[23] L. Macchion, R. Fornasiero, A. Vinelli, Supply chain configurations: a model to evaluate performance in customised productions, Int. J. Prod. Res. 55 (5) (2017) 1386–1399.
[24] L.M.S. Dias, A.A.C. Vieira, G.A.B. Pereira, J.A. Oliveira, Discrete simulation software ranking — a top list of the worldwide most popular and used tools, 2016 Winter Simulation Conference (WSC), 2016, pp. 1060–1071.
[25] K. Sahoo, S. Mani, GIS based discrete event modeling and simulation of biomass supply chain, Proceedings - Winter Simulation Conference, 2016, pp. 967–978.
[26] B. Ponte, E. Sierra, D. de la Fuente, J. Lozano, Exploring the interaction of inventory policies across the supply chain: an agent-based approach, Comput. Oper. Res. 78 (2017) 335–348.
[27] M. Golfarelli, S. Rizzi, Data Warehouse design: Modern principles and Methodologies 5 McGraw-Hill, New York, 2009.
[28] R. Elmasri, Fundamentals of Database Systems, Pearson Education, India, 2008.
[29] E. Costa, C. Costa, M.Y. Santos, Efficient big data modelling and organization for Hadoop Hive-based data warehouses, European, Mediterranean, and Middle Eastern Conference on Information Systems, EMCIS 2017, Lecture Notes in Business Information Processing, 299 2017, pp. 3–16.
[30] C. Costa, M.Y. Santos, The suscity Big Data warehousing approach for smart cities, ACM International Conference Proceeding Series, Part F1294 2017, pp. 264–273.
[31] J.R. Lourenço, V. Abramova, M. Vieira, B. Cabral, and J. Bernardino, “NoSQL Databasesdatabases: aA software engineering perspective,” 2015, pp. 741–750.
[32] P. Grover, A.K. Kar, Big Data analytics: a review on theoretical contributions and tools used in literature, Glob. J. Flex. Syst. Manag. 18 (3) (2017) 203–229.
[33] S. Mohanty, M. Jagadeesh, H. Srivatsa, Big Data imperatives: Enterprise ‘Big Data’ Warehouse,‘BI'implementations and Analytics, Apress, 2013.
[34] R.G. Goss, K. Veeramuthu, Heading towards Big Data building a better data warehouse for more data, more speed, and more users, ASMC 2013 SEMI Advanced Semiconductor Manufacturing Conference, 2013, pp. 220–225.
[35] M.Y. Santos, et al., A Big Data system supporting Bosch Braga Industry 4.0 strategy, Int. J. Inf. Manage. 37 (6) (2017) 750–760.
[36] N. Nodarakis, S. Sioutas, A. Tsakalidis, G. Tzimas, Using Hadoop for large scale analysis on Twitter: a technical report, arXiv Prepr. arXiv1602.01248 (2016).
[37] R.S. Kv, N.P. Kavya, Trend analysis of e-commerce data using Hadoop ecosystem, Int. J. Comput. Appl. 147 (6) (2016) 1–5.
[38] A. Thusoo, et al., Hive - a petabyte scale data warehouse using Hadoop, 2010 IEEE 26th International Conference on Data Engineering (ICDE 2010), 2010, pp. 996–1005.
[39] A. Thusoo, et al., Data warehousing and analytics infrastructure at Facebook, Proceedings of the 2010 international conference on Management of data - SIGMOD ’10, 2010, p. 1013.
[40] F. Di Tria, E. Lefons, F. Tangorra, Design process for Big Data warehouses, DSAA 2014 - Proceedings of the 2014 IEEE International Conference on Data Science and Advanced Analytics, 2014, pp. 512–518.
[41] C. Costa, M.Y. Santos, Big Data: state-of-the-art concepts, techniques, technologies, modeling approaches and research challenges, IAENG Int. J. Comput. Sci. 44 (3) (2017) 285–301.
[42] D. Simchi-Levi, et al., Identifying risks and mitigating disruptions in the automotive supply chain, Interfaces (Providence) 45 (5) (2015) 375–390.
[43] S.A. Masoud, S.J. Mason, Integrated cost optimization in a two-stage, automotive supply chain, Comput. Oper. Res. 67 (2016) 1–11.
[44] J.-H. Thun, D. Hoenig, An empirical analysis of supply chain risk management in the German automotive industry, Int. J. Prod. Econ. 131 (1) (2011) 242–249.
[45] O. Kırılmaz, S. Erol, A proactive approach to supply chain risk management: Shifting orders among suppliers to mitigate the supply side risks, J. Purch. Supply Manag. 23 (1) (2017) 54–65.