DOI: 10.15439/2014F472
ACSIS, Vol. 3

# Research Problems Associated with Big Data Utilization in Logistics and Supply Chains Design and Management

Silva Robak
Uniwersytet Zielonogórski, ul. prof. Z. Szafrana 4a, 65-516 Zielona Góra, Poland
Email: s.robak@wmie.uz.zgora.pl

Bogdan Franczyk
Uniwersytet Ekonomiczny we Wrocławiu, ul. Komandorska 118/120, 53-345 Wroclaw, Universität Leipzig, Germany
Email: franczyk@wifa.uni-leipzig.de, bogdan.franczyk@ue.wroc.pl

Marcin Robak
XLogics Sp. z o.o., ul. Kostrzyńska 4, 65-127 Zielona Góra, Poland
Email: m.robak@xlogics.eu, Uniwersytet Zielonogórski,WEIiT, m.robak@weit.uz.zgora.pl

***Abstract*—One of contemporary big challenges in information systems are the issues associated with coping with and utilization of the vast amounts of data. In this position paper we present few research problems emerging in association with occurrence of big data, with focus on domain of logistics and supply chains. We also reveal a number of influence factors on supply chains design and management which are related to big data. We show that the association of data science and predictive analysis with domain knowledge can build an important improvement for logistics and supply chain design and management. We point out how the domain problems in the key managerial business components could be solved better with big data applications for the stakeholders involved in the supply chains. We also highlight open problems in this domain and chances for the future.**

## I. Introduction

IN THE contemporary world logistics companies have to face unprecedented challenges. As a result of globalization the amount of data arising in supply chains is raising, the competition is becoming fiercer and the customers often expect integrated services, what requires a close cooperation between involved organizations. The companies have to adapt to new business models and rethink their role and position in their value chain regarding the potential possibilities given by the utilization of big data to add value for their customers and suppliers. This requires changes from logistic companies in their way of thinking about the supply chain design and management, and at the same time in their information technology view to support the collaborative decision making.

The problem of the appropriate information technology environments for collaborative processes between business participants is still present, since appropriate IT infrastructures for utilization of big data are needed. There are data 'siloes' from diverse applications like ERP or customer relationship management systems (CRM), etc. and the collaboration between business partners may require undertaking steps for IT environment integration, such as one of the known enterprise application integration solutions or usage of the Web services [1].

In our previous papers we approached the problems of advantageous utilization of vast amounts of big data in supply chains and also the information integration issues in order to overcome the data silo problem. The proposal of possible utilization of the Linked Data [2] as an integration solution for business process management BPM in supply chains networks we have already presented in [3]. In [4] we investigated the appropriate IT architectures for big data used in association of cloud computing facilities [5] and the utilization of common (open stated) data format as it is offered by Linked Data for data silos integration purposes. In this position paper we recommend the utilization of big data in conjunction with Data Science and Predictive Analysis, as appropriate for logistics with emphasis on the value-adding partnerships in supply chains.

A supply chain is defined as a network that comprehends all the organizations and activities associated with the flow and transformation of goods, starting from raw material stage through the whole process, to the end user, as well as the associated information flow [6].

In the inter-organizational information systems, which link companies to their suppliers, distributors and customers, a movement of information through electronic links takes place across organizational boundaries between separately owned organizations. It requires not only electronic linkage in form of basic electronic data interchange systems, but also interactions between complex cash management systems or by accessing shared technical databases. The problems with sharing and exchange and also utilization of information are viable in supply chains contexts.

A business process in a supply chain consists of one or more than one related activities that combined together respond to the need for a business action. The processing steps in a workflow might go through numerous data transformations (geographic, technological, linguistic, syntactical and semantic transformations) [6]. We will show that data science and predictive analysis may leverage the utilization of big data for business processes in supply chains in business networks. For this aim the rest of the paper is organized as follows.

In Section 2 we explain the V-model and the key characteristics of big data and its storage formats appearing in supply chains. In Section 3 we reveal the concepts of predictive analysis and data science. We examine the possible added value resulting from their application for big data in supply chain design and management. In Section 4 we point out some open research problems in logistics and supply chain design management. In the last Section of this position paper we conclude our work.

## II. BIG DATA IN LOGISTICS AND SCM

### *A. Big Data V-Model*

Big data is referred to data that goes beyond the processing capacity of the conventional database systems. In addition to the aspect that it is big (e.g. a huge number of small transactions, or continuous data streams from sensors, mobile devices etc.) it may move too fast, or does not fit the structure of traditional (i.e. relational) database architectures. Big data also may have a low value for further usage before processing it [7].

When we denote a big amount of data as "big data" it has to cover the 3V model with three basic features such as: volume, velocity and variety [7].

Volume of big data denotes its massive character, i.e. a huge amount of information involved. The big volume of data can be beneficial for the data analysis aims. It may improve the analytics models by having more cases available for forecasts and increase the number of factors to be considered in the models making them more accurate. On the other hand, the volume bears potential challenge for IT infrastructures to deal with big amounts of data, especially when taking into account its second V-feature – velocity.

The velocity in which data flows into organization or the expected response time to the data is the second V-feature of big data. Big data may arrive quickly - in (near) real-time (i.e., near-time). If data arrives too quickly the IT infrastructures may be not able to respond timely to it, or even to store all of it. Such situations may lead to data inconsistencies.

The variety of big data comes from its many potentially different data sources. Therefore big data may have diverse structures and forms, not falling into the rigid relational structures of SQL databases without loss of information. Some of data may be saved as blobs inside traditional data bases. The IT infrastructures for big data are denoted as NoSQL (i.e., "not only SQL" [8]). Examples of diverse sources and kinds of data are standard business documents, transactional records and unstructured data in form of images, recordings, HTML documents, Web pages, text and e-mail messages, streams from meters and environment sensors, GPS tracks, click streams from Web queries, social media updates, data streams from machines' communication or wearable computing sensors, etc. [4].

In accordance with authors we add the fourth V-feature for big data, which is its value. The big data value V-feature denotes the need for processing it before using it in order to make it valuable for analysis purposes ([8] and [9]).

We claim that the fourth V-feature bears a special importance for logistics and supply chain design and management.

The big scale usage of available and generated data is made possible for organizations owing to cloud computing paradigms, such as Infrastructure as a Service IaaS, Storage as a Service SaaS, which revolutionized the way the computing infrastructures are used [5].

### *B. Big Data Characteristics (Data structures)*

In the beginning of this Section the V-model for big data with its four characteristic V-features: volume, velocity, variety and value have been considered. Based on it we now define big data as:

"data whose scale, distribution, diversity, and/or timeliness require the use of new technical architectures and analytics to enable insights that unlock new sources of business value." [10].

The variety of data streams in logistics and SCM may be regarded as structured, semi-structured, "quasi"-structured or unstructured. The structured data contains defined data types, formats and structures. Transaction data from Online Transaction Processing (OLTP) and Online Analytical Processing (OLAP) can be named as example.

The semi-structured data is represented by textual data files with discernable patterns, what enables data parsing. The examples of semi-structured data are self-describing XML data files defined by XML Schema.

The quasi-structured are irregular data formats that can be formatted, but only with additional effort, tools and time. An example of such data could be Web click-stream data.

As the unstructured we denote data with no inherent structure that can be stored in different types of files. The fact that part of data is unstructured, or rather, it lacks a structure appropriate for storage in conventional SQL databases, implies that other solutions are needed.

The conventional IT infrastructures in supply chains include structured data in form of OLTP and OLAP systems. The traditional OLTP systems support the transactional systems with highly structured SQL databases, whereas the OLAP systems contain aggregated historical data in form of cubes. The OLTP systems deliver simple reports, while OLAP systems (Data Warehouses) are suited for (traditional) business intelligence applications with reporting facilities on business statistics, performance, etc. on the basis of structured (analytical) historical data. On the opposite of partially low value big data the OLAP and OLTP provide only high quality data.

As stated above these both databases forms are unsuitable for all big data formats. The data stored there has to have fixed structure, which is conflicting with variety of big data. Hence the analytical OLAP systems contain only historical data; they are not suitable for big data applications. Another problematic issue may be due to the velocity of big data.

As the rigid SQL data structures are insufficient for big data applications, other OLT solutions in form of 'key-value stores' like "NoSQL OLTP" MongoDB, AmazonDynamo or Windows Azure Table Storage can be used [11].

For storing and analyzing massive data sets the "NoSQL warehousing" open-source Apache Hadoop [11] can be applied. It is a highly distributed and fault-tolerant framework for software development with its own highly distributed HDFS file system, and a MapReduce framework for writing and executing distributed algorithms.

The conventional IT structures may encounter problems with storing variety of data and immediately reacting to it. It is because of big data amounts on (also unstructured) data arriving in near-time. It is apparent that efficient dealing with big data in logistics requires for new data architectures, analytics sandboxes and tools. The IT infrastructure for the data platform will be preferably cloud-based.

## III. Predictive Analysis and Data Science

In this Section we present the special role of data science and predictive analysis in logistics and supply chain development and management.

Data Science is the application of quantitative and qualitative methods to solve the relevant problems and predict outcomes. Its importance is due to the fact that with growing amounts of data the domain knowledge (here logistics) and the generalizable extraction of knowledge from data (data analytics) cannot be separated. Thus, the desirable professionals involved in supply chain design and management should posses both kinds of skills – the analytical skills and at the same time a deep understanding of the business domain and its management.

There are several roles to be covered by the practitioner of data science like: the analysts, the data professionals and also the technology and data enablers. Data scientists in the role of data analysts should possess a deep analytical talent and also an advanced training in quantitative disciplines like mathematics, statistics and machine learning.

The data savvy professionals (such as analysts and managers) are preferably people with basic knowledge of statistics and advantageously also machine learning, who should be able to define key questions that can be answered by using advanced analytics. The technology and data enablers are data scientists providing technical expertise needed to support analytical projects; their skills sets include computer programming and database administration knowledge.

The key activities of data scientists in logistics include providing services to other stakeholders like data engineers, data analysts, business analysts and the users in a line of business. Such activities comprise the reframing business challenges as analytics challenges, design, implementation and deployment of statistical models and also the data mining techniques, especially for big data. A crucial aspect in data science is thereby included in creating insights leading to actionable recommendations to help business to gain a competitive edge [12].

For supply chain design and management an advantageous skill set in the data science includes various skills, where each may have a different importance (weight). The disciplines to be covered are such as: statistics, forecasting, optimization and discrete event simulation, applied probability, analytical mathematical modeling, finance, economics, marketing, and accounting. For all the above mentioned disciplines, due to big data utilization, the focus will be different than in a traditional approach. More important become skills in conjunction with broad awareness of many different methods in comparison to the classical point of view. For instance, for the aims of forecasting the understanding the application of qualitative and quantitative methods will become more important than the understanding of the underlying stochastic processes.

Predictive analysis is a subset of data science. It encompasses a variety of statistical techniques that enable to take advantage of the patterns found in historical and transactional data. In logistics it could help to optimize business operations, to identify business risks (and security aspects), to predict new business opportunities and to fulfill the law or regulatory requirements. The business value of predictive analysis (data science) and data mining will be higher than gained from conventional business intelligence due to optimization, predictive modeling, forecasting and analysis of vast data resources (big data).

In logistics the predictive analysis uses both quantitative and qualitative methods to estimate the past and future behavior of the flow and storage of inventory, as well as the associated costs and service levels. On the other hand, the SCM predictive analysis also uses both quantitative and qualitative methods to improve supply chain design and competitiveness by estimating past and future levels of integration of business processes among functions or companies, as well as the associated costs and service levels. The value of predictive analysis is not to be scoffed at. Together with the appropriate analytics tools it may become a decisive competitive asset.

## IV. Research Problems in Logistics and SCM

The open research problems in logistics and supply chain management can be viewed from the perspective of key managerial components of business logistics and on the other hand the different category of the stakeholder. The main business functions are such as forecasting, inventory management, transportation management and transportation and human resources. The stakeholders (user) such as carrier, manufacturer, retailer may be examined how they could benefit from usage of big data all the business functions in logistics and supply chain management.

For instance, considering the forecasting, the user carrier, by relating to forecasting, could predict a out the time delivery, etc. The manufacturer would be able to make an early response to extremely negative or positive customer sentiments, etc. The inventory manager may plan capacity availability in real time, etc. For the retailer it could mean an improvement in perpetual inventory system accuracy.

The similar approach may be conducted for the type of data and the management functions: inventory, transportation, customer and supplier relationship management to identify the problems to be solved with application of predictive analysis by using big data. Further consideration to the evolution of logistics [14] due to internet-based applications can be found in [15], [16] and [17].

## V. Conclusion

This position paper has provided a research perspective on contemporary problem and chances in the domain of logistics, supply chain design and management in conjunction with data science, predictive analysis and big data. We have highlighted the various aspects of big data and data management in logistics and supply chains and pointed out how to efficiently use big data across the supply chains development and management.

We have categorized the potential applications components of big data in four functional categories: forecasting, inventory management, transportation management, and human resources and proposed activities based on big data for three categories of user (carrier, manufacturer, retailer). The insights won from individual components should be integrated into one global strategy on usage of predictive analysis and data science in logistics domain.

The discussion presented in this paper highlights the fact that it is important to consider using big data and predictive analysis and data science along with particular domain knowledge concerns for achieving enhanced results throughout a supply chain life cycle.

As we stated in [4], in the past vendors had to exploit earlier period's structured data in order to analyze stockholders needs and sentiments and to increase the optimal performance and the potential business value. We have examined the nowadays common solutions for data storage options for the decision making support. With the opportunity of usage of big data and cloud computing technologies, along with raising all-embracing connectivity with involved stakeholders in supply chains networks, results in the possibility of accessing current data in near-time and in getting a near-time feedback. This bears legitimate chances for almost immediate improvement of the relationships with the supply chain's stakeholders and therefore increases the agility and ability to react in real-time to the environmental changes.

One of the advantages of predictive analysis and data science is that they may provide better insights than it would be possible with traditional business intelligence systems. Thus, the high quality decision support becomes attainable, and usage of predictive analysis and data science methods will enable extending its application in decision making.

In this position paper we have presented an introduction in big data usage in logistics and supply chains. We expect that the application of predictive analysis and data science in this domain will shift the emphasis from traditional point of view of relevant aspects in many comparative disciplines like statistics, forecasting, optimization, discrete event simulation, applied probability, data mining, etc. The main criteria will become the speed of decision making, throughput and analysis flexibility.

In the future work further aspects like economic evaluation of applying big data and linked data concepts in supply chain management should be considered, also from the point of view of management theory.

## References


[1] Web Services Activity, W3C Working Group, http://www.w3.org/ws

[2] W3C LinkedData, 2011, www.w3.org/wiki/LinkedData

[3] S. Robak, B. Franczyk, and M. Robak, Applying Linked Data concepts in BPM, *Proceedings of the Federated Conference on Computer Science and Information Systems FedCSIS*. Wrocław 2012. IEEE Conference Publications, ISBN: 978-1-4673-0708-6, pp. 1105-1110.

[4] S. Robak, B. Franczyk, and M. Robak, Applying Big Data and Linked Data concepts in Supply Chain management, *Proceedings of the Federated Conference on Computer Science and Information Systems FedCSIS*. Kraków 2013. IEEE Conference Publications, pp. 1215-1221.

[5] D. Agrawal, S. Das and A. E. Abbadi, „Big data and cloud computing: current state and future opportunities". *EDBT 2011*, March 22-24, 2011, Uppsala, Sweden. ACM 978-1-4503-0528-0/11/0003, DOI http://dx.doi.org/10.1145/1951365.1951432.

[6] M. P. Papazoglou, and P. M. A Ribbes, *E-business: organizational and technical foundations, John Wiley and sons*. London 2006, pp.88-90.

[7] E. Dumbill, "*What is big data? An introduction to the big data landscape*", Strata O'Reilly, 11 January 2012, http://strata.oreilly.com/2012/01/what-is-big-data.html

[8] S. Wrobel, „*Big Data – Vorsprung durch Wissen*", Fraunhofer-Institut für Intelligente Analyse- und Informationsverarbeitungsysteme IAIS. Presentation, www.iais.fraunhofer.de, 2012.

[9] I. Mitchell and M. Wilson, "Linked Data. Connecting and exploiting big data", Fujitsu Services Limited, March 2012, www.fujitsu.com.uk.

[10] B. Franczyk, "*Data science, predictive analysis and big data in logistic, SC design und management"*. Universität Leipzig, Wirtschaftswissenschaftliche Fakultät. Institut fur Wirtschaftsinformatik, Presentation 2014.

[11] The Apache Hadoop Project. http://hadoop.apache.org/core/ , 2009.

[12] La Ponsie, Maryalene, "*Data scientiscs: the hottest job you haven't heard of*" online http://jobs.aol.com/articles/2011/08/10/data-scientist-the-hottest-job-you-havent-heard-of/

[13] D. C. Luckham, *Event processing for business: organizing the real-time enterprise*. Hoboken, New Jersey: John Wiley & Sons, Inc., p.3. 2012.

[14] H. Baumgarten, *Das beste der Logistik*. Springer Verlag, Berlin 2008.

[15] S. Chopra, and P. Meindl, "*Supply chain management: strategy planning, and operation*", 3rd ed., Prentice Hall, 2007, pp.427.

[16] A. Matopoulos, and E. -M. Papadopoulou, "The evolution of logistics service providers and the role of Internet-based applications in facilitating global operations" in *Enterprise Networks and Logistics for Agile Manufacturing*, L. Wang, and S.C.L. Koh, Eds., Springer, 2007, pp. 298-304, DOI 10.1007/978-1-84996-244-5_14.

[17] R. Sethuraman and S. K. Kundharaju, "Top 7 Tips for Big data to optimize Supply Chains. 5 Februar 2013, http://risnews.edgl.com/retail-trends/Top-7-Tips-for-Utilizing-Big-Data-to-Optimize-Supply-Chains86163.