Big data processing in the logistics industry

Snezhana SULOVA[1]

[1] University of Economics, Varna, Bulgaria
ssulova@ue-varna.bg

Abstract. The aim of modern logistics is to achieve maximum connectivity in the supply chain. Companies are using increasingly innovative technological solutions, which creates the opportunity of generating a wide variety of data. This leads to several challenges and the need to change data storage and processing models. The aim of the study is to analyze the technological aspects of the digital transformation in logistics and to propose a conceptual framework for big data management and processing in the logistics industry. It is based on the discovery of existing prototype methodologies for big data processing which are used in all areas of business, as well as on the research of existing specific approaches to the processing of different types of big data in logistics. Basic principles for building a modern architecture for managing and processing big data in logistics are presented. The defined framework can be used by the companies to process structured, semi-structured and unstructured data in real time or for batch processing and to help optimize several business processes in the logistics industry. As a result, using it will help the analytical processes in these companies and it will be possible to make informed business decisions in dynamic conditions and in globalization. A software implementation of a conceptual framework with the Apache Handoop open-source software is proposed. The study is part of Project BG05M2OP001-1.002-0002-C02 "Digitalization of Economy in a Big Data Environment"

Key words: logistics, big data, digitization, conceptual framework.

1. Introduction

Developments in the modern economy are mainly due to the process of digitalization. According to current research, 87% of small companies and 6% of large companies in Western Europe have invested in the digitalization and automation of their production (pwc.de, 2017). In the field of logistics, old ways of distributing physical goods have been replaced with modern solutions, which are based upon the application of innovative technologies that have been integrated into the existing supply chain management (SCM) or have created completely new working concepts.

The use of innovative digital solutions, with the goal of optimizing and changing the functioning model of business, creates enormous prospects for businesses to develop, but at the same time, the new models of work also bring with it many challenges.

Modernization, digitization, and cloud computing have resulted in the appearance of ever greater volumes of data. The creation and use of data is growing rapidly throughout the world. In 2020, data volume was 59 zettabytes, and it is expected to be 149 zettabytes by 2024 (statista.com, 2021). On the one hand, this creates more opportunities to improve business intelligence analyses, which could be provided in real-time allowing for informed decisions to be made, but on the other hand, it will create a need to change management models and how organizations process data.

The goal of the current research is to analyze the technological aspects of the digital transformation of logistics, by focusing on the increasing volumes of data and to propose a conceptual framework for the management and processing of data in the logistics industry.

2. Technological aspects of the digital transformation of logistics

The initiative “Industry 4.0” began with the acceptance of a high-tech strategy for the development of industry in Germany (Lydon, 2014). Following that decision, many studies have analyzed the content, scope and technology that led to a fourth industrial revolution. Kondratiev (2019) points out that the revolution is due to: the use of integrated sensors through which devices are able to communicate with each other or better known as the Internet of Things (IoT) technology; working in real-time with the goal of optimizing expenses and the

quality of production through the application of Big Data Analytics (BDA); autonomous robots and 3-D printers. Another comprehensive study of the main technological trends of Industry 4.0 identifies: virtual and augmented reality (VR, AR); cloud computing; robotization; BDA; cyber-physical system (CPS); cybersecurity; Internet of People (IoP), Internet of Services (IoS); Industrial Internet of Things (IIoT); semantic technologies; simulation and modeling (Ghobakhloo, 2020).

The processes of digital transformation in logistics led to the emergence of the Logistics 4.0 concept. The term, Logistics 4.0, means the specific application of Industry 4.0 to the field of logistics (Raksoy, Koçhan, Ali, 2020, p. 21).

The range of technologies, which are the basis for the transformation of traditional organizations into smart businesses, are constantly growing (Lazarova, 2019). It is necessary to note that in order to create a digitally connected world, a complete change in business models is needed. Turchi (2018) has developed a framework for the digital transformation of a business, which is called “The Pyramid of Digital Transformation”. He defines 3 levels: technologies, new business models and strategies for development (Fig. 1).

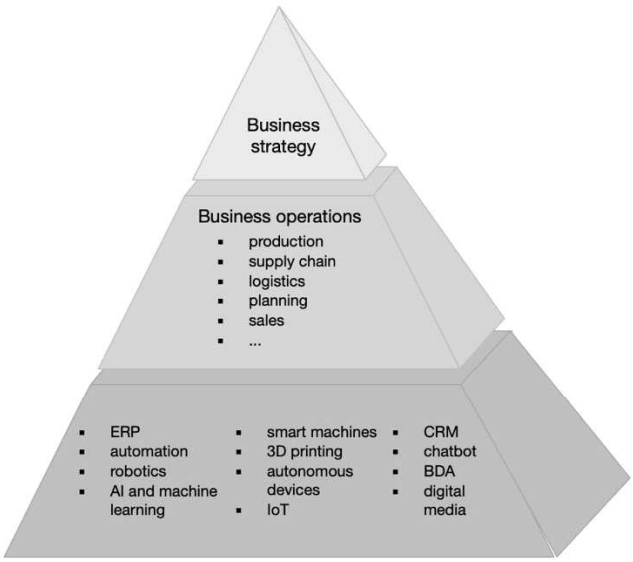

Figure 1. Pyramid of Digital Transformation
Source: (Turchi, 2018)

Each of the three levels of the pyramid affects each of the other elements of the framework and therefore, a successful digital transformation must use an integrated method and must create new strategies through Industry 4.0.

The focus of our research is the technological changes, which form the necessary base for the application of modern concepts to logistics activities. The key technological changes are**: vertical and horizontal system integration; Internet of Things; cloud computing; blockchain; autonomous robots; virtual and augmented reality** (Fig. 2).

The processes of digital transformation of logistics requires system connectivity and the presence of vertical integration as is related to the systems of the organization and to what extent they are integrated with each other as well as the horizontal integration with business partners. The creation of large, distributed, complex systems, which work flexibly as a self-organized structure using a new type of production facility are the foundation of **cyber-physical systems**. CPS ensures flexible mass production on demand and flexibility in the quantities produced (Rojko, 2017). Some authors point out that CPS has the following main characteristics: heterogeneous character; connectivity and networking abilities; as well as their ability to work based on the “Software as a Service” model; modularity; autonomy; decentralization; real-time operation; powerful computing abilities; dynamic possibilities for reconfiguration and adaptability (Napoleone, Macchi, Pozzetti, 2020).

With the development of CPS technologies over the past few years, physical devices, which use computers and networks to expand their functions, have led to modernized industrial products and technologies and have significantly improved their ability to compete in major industrial sectors (Zhang, D. et al., 2021)

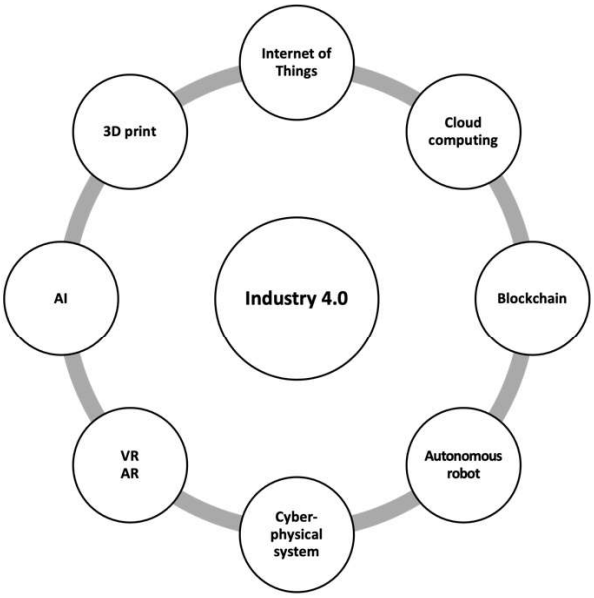

Figure 2. Key components for the digital transformation of logistics (Logistica 4.0)
Source: Own elaboration

There is a direct link between CPS and the **IoT** concept, whose main function is the collection of data from various physical entities using devices that have access to the internet. IoT “proposes the construction of communities in which the objects around us - often called ‘intelligent objects’ – can be connected through the Internet” (Stoyanov and Popchev, 2017). Thanks to these technologies, the concepts of smart production, smart buildings, automobiles, and logistics have been developed. Smart logistics pertains to the precision, reliability and effectiveness of logistics performance due to improved information and the use of data (Chaopaisarn & Woschank, 2020). In the field of logistics, smart logistics is related to real-time work, working conditions and the maintenance history of the production equipment. In the field of logistics, IoT can track the routes and alert to any deviations from the planned route, analyze the location of vehicles and route data and upon discovering difficulties in the supply chain, such as a traffic jam, it can generate alternative routes to optimize the supply chain.

Statistical data shows that at the end of 2018 22 billion IoT devices had been used throughout the world, and the data predicts that by 2030 there will be about 50 billion creating an extensive network of connected devices (Statista.com, 2021).

The development of the IoT concept has directly contributed to an increase in data volume coming from various sensors, GPS location tracking devices, radio frequency identification (RFID) tags, and others. This has led to an increase in demand for computer (computing) resources for storing and processing information. The cloud-based business model provides an opportunity for small and medium-sized businesses to meet the challenge of managing ICT infrastructure, platforms and services (mi.government.bg, 2017). Data shows that in 2020 more than 1/3 of EU businesses used cloud computing mainly for email services and file storage (Eurostat, 2021).

Technological factors that stimulate the development of cloud computing are: distributed processing; wide area networks and internet; high-performance hardware and virtualization (Emilova, 2016). Cloud computing can be found in every sector of the economy due to its advantages, such as easy access to shared resources, optimization of expenses, flexibility, and virtualization of resources.

Blockchain technology also plays a leading role in the process of digitizing logistics. The Gartner Research and Consulting company pointed out that blockchain was one of the ten most important strategic technologies for 2020 (Gartner, 2020). Blockchain is connected to the development of decentralized applications for the automation of business processes and data processing (Filipova, 2018, p. 83). It can be utilized in financial and insurance fields (Kirov, 2020), for “smart contracts”, for retail operations and the supply chain of products, goods and services (Dimitrova and Semova, 2018), property management (Stoyanova, Vasilev, Cristescu, 2021) as well as others.

Robots are used in many industries for completing complex tasks. The increased use of industrial robots by companies has accelerated the process of digital transformation. Intelligent technologies provide devices with new qualities and possibilities. Robotized processes can be found in the treatment of chemicals, the production of pharmaceuticals, food and drinks, and in logistics (Proctor and Wilkins, 2019).

Augmented virtual reality also contributes to the formation of an intelligent business. AR combines the physical and the real world with the computer world of generated data, with the goal of turning a virtual object into reality. **3D printing** (additive manufacturing) is important for some industries whose work necessitates the building of different complicated geometric constructions. In the past few years, larger investments have been made in 3D printing and studies show that by 2023 these investments will reach $32,78 billion (Proctor and Wilkins, 2019). Important benefits of 3D printing include decreasing the weight of components and waste reduction.

All of the technologies mentioned, combined with the possibilities, which intelligent technologies have for adapting their behavior and for autonomous work on the basis of the analysis of results from previous actions is fundamental for the formation of a technological base for the processes in digital transformation.

To summarize, it can be noted that the processes of greater interconnection between information systems, the presence of dynamic activities, the market conditions, the use of constantly enriching number of technological solutions in logistics has led to the generation, storage and processing of large volumes of data.

3. Data as a condition of digital transformation in logistics

The word “data” comes from the Latin word dare, which means “something, which is given” - observation or fact about a given object (Kotu and Deshpande, 2019, p. 39). Data is a vital and irreplaceable commodity for every organization. Today businesses work with operational data, as well as with large volumes of data, which are generated with the help of internet applications; transactions, which happen in real-time; publications in social media; pictures; mobile phone messages, etc. This necessitated the emergence of the concept “Big Data”.

Big data is a collection of chosen data, whose volume, speed or variety are so large that it is difficult to store, manage, process and analyze the data using traditional databases and instruments for data processing (Bahga and Madisetti, 2019, p. 25). The Gartner consulting company has defined three main characteristics of big data known as the so-called “3V” model:

- Volume. Data volume can be distinguished as the most essential characteristic of big data. As it has already been pointed out, all the digital transformation processes lead to the generation of large volumes of new data. Internet, social networks, use of mobile devices, measuring devices are all capable of generating large volumes of data.
- Velocity. Currently, data appears extremely fast. With the increased use of sensors and devices working with IoT, data is generated and transferred with great speed, which creates the need to process it in a timely manner, in real-time mode;
- Variety. Big data has a heterogeneous character. Data is in different formats - operational data from databases, text documents, emails, videos, audio files, sensor data, etc.

Some authors (Sharda, Delen, Turban, 2020, p. 125) divide data into three different types depending upon the data structure as it can be seen in Fig. 3.

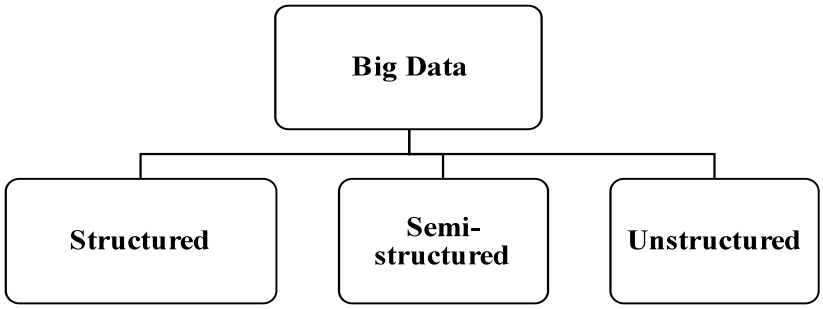

Figure 3. Types of Big Data
Source: (Sharda, Delen, Turban, 2020, p. 125)

Structured data has a form, which is suitable for computer processing. Often the data is stored and processed by Database Management System (DBMS). The data is generated by ERP, CRM or other business systems, sensor devices working with IoT technology, and surveys. **Semi-structured data** as a whole has some sort of organizational structure that eases its ability to be processed. It includes tags and elements (metadata), which are used to group and describe them. Typical examples are JSON, XML and CVS files.

Unstructured data does not have a specific structure or model of organization. It is in the form of text files, email messages, devices for video and audio recordings. According to prevailing assessments, unstructured data constitutes 80-90% of all data (Cio.com, 2019).

Other authors identify data according to its source and categorize it as: social media data; smart and IoT devices data, sensor data and transaction data (Storeya & Song 2017).

In order to better define the characteristics of big data including the challenges faced in processing big data, seven characteristics have been added to the three main ones. Firican (2017) uses the “10V” model to describe big data. In table 1, the seven additional distinguishing features of big data are given.

Table 1.

Characteristics of big data

Characteristics	Description
Value	It shows that with proper processing and analysis of big data, significant benefits can be obtained.
Variability	Data has many dimensions due to its varying types and sources.
Veracity	Data must have a reliable source and origin in order to be relevant for analysis.
Validity	Data must be accurate and true.
Vulnerability	Describes how long data that has been obtained from the original source has been available and how long it should be stored.
Volatility	Describes how long the obtained data is and how long it should be stored.
Visualization	Shows how the data processing results are presented.

Source: Own elaboration

With an aim of characterizing big data in greater detail and defining the challenges connected with its storage and processing, Shafer (2017) has presented an updated list with 42 V’s.

Big data from logistics is generated by the following sources:

- **Traditional computer systems**, which maintain the logistics business processes. They are: Enterprise Resource Planning systems (ERP), Warehouse Management Systems (WMS), Transportation Management Systems (TMS), Customer Relationship Management systems (CRM), Supply Chain Management systems (SCM) and others. A detailed overview of information systems used by modern logistics as well as the volume of information in supply chains has been made by Vasilev (2017).
- **Data from Geographic Information Systems** (GIS), which has a rather diverse nature - numerical values, text, images (pictures), graphics, audio, animation and video (Gergova et al., 2017). Global Positioning System (GPS) is of great importance to transport and logistics companies, as it allows for the development and application of spatial and temporal models for the optimization of the routes.
- **Data from devices using IoT** concepts. IoT systems work with a variety of large volumes of data, which is collected from sensors, RFID and mobile devices. Internet of things is one of information technologies which can be used in 5PL logistics to collect and analyze large quantities of data (Zaychenko et. al, 2021). The data is most often found in the XML format (eXtensible Markup Language), JSON (Javascript Object Notation), PNG (Portable Network Graphics), CSV (Comma-separated values), XDR (eXternal Data Representation), RDF (Resource Description Framework);
- **Data from Augmented Reality** technology applications, which can be integrated into vehicles and provide a video stream of data and computer-generated graphics.
- **Company websites, web and mobile applications, social media profiles**. Web resources, sites, web applications and mobile applications are sources of various data. On the one hand, when one interacts with them usable data is generated, and on the other hand, these sources, themselves, have their own content and they also allow the creation of new content, which can also be a source of data. Currently, a tendency is being observed that traffic from wireless and mobile devices is quickly increasing and this generates new data for the analysis of users’ behavior.

- **Text documents and email messages**. Text documents with different structures and formats are used and exchanged in the logistics industry. Transforming these documents using natural language processing (NLP), they can become a useful source of knowledge.
- **Data, collected by governments through open databases**, such as those connected to intellectual property, civil infrastructure, scientific developments, and protection of the environment.
- **Data, connected with government regulation –** patents, regulatory activities, tax information.
- **Other data** from accompanying activities, such as data from advertising campaigns for example.

The importance of big data is since analyses can be performed on it which an organization will value. Many businesses are currently unable to process the large volume of data that they possess, or they are unable to cope with the speed with which new data is produced by sensors and social networks. Although traditional relational data management systems ensure speed in processing transactions, they do not have the necessary scalability to work with semi-structured and unstructured data. The digital environment, as well as the peculiarity and variety of work with large volumes of data requires the use of specific and improved new approaches for the processing and storage of the data. Several technologies exist for the storage, processing and analysis of big data. MapReduce, Hadoop and NoSQL have been highlighted by analysts as the most important technologies for digital transformation (Sharda, Delen, Turban, 2020, p. 523).

MapReduce is a computing technology, which can be implemented on many systems. “It can be viewed as a programmed data-processing module, whose purpose is to compact large quantities of information into useful summary results” (Petrov and Trifonova, 2019, p. 127). MapReduce performs two tasks: Map and Reduce. The first phase is to process inputted data. In order to complete this phase, one of the computers is assigned to be the main node and it receives the inputted data, divides it into parts and sends it to other computers (working nodes) for initial processing. In the Reduce phase, previously processed data is reduced. The main node organizes the results - the solution to the problem, which was initially formulated- based upon the answers received from the working nodes (Fig. 4).

High performance is reached when the process is reduced to small units of work, which can be completed in parallel across hundreds, potentially thousands, of cluster nodes. This process helps in processing and analyzing large volumes of data. MapReduce is available both commercially and for libraries with open code providing a wide spectrum of analytical possibilities. Apache Mahout, for example, is a library for open source machine learning, which includes cluster algorithms, classification and it works with MapReduce.

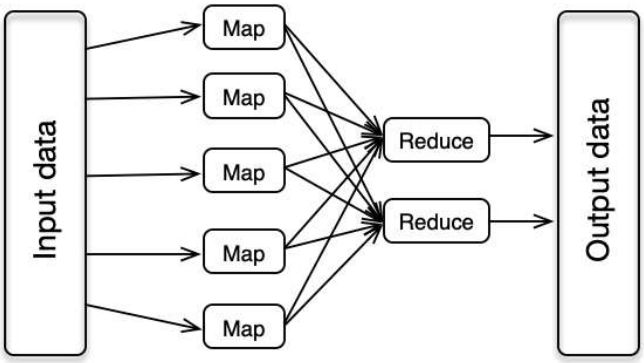

Figure 4. MapReduce Technology
Source: Own elaboration

Hadoop is a software framework with open code for processing, storing and analyzing large quantities of distributed unstructured data. It was created by Doug Cutting, Mike Cafarella and is now a software project of the Apache Foundation, who are continuously working to perfect it. Hadoop has two main components: Hadoop Distributed File System – HDFS and an environment for MapReduce inquiries. The client sends a Map query written in Java to one of the cluster nodes (Job Tracker), it initiates, coordinates the tasks and sends a query to the respective nodes. Processing happens on each node simultaneously or in parallel.

The Hadoop software instrument package includes NoSQL databases like Cassandra and HBase, which are also used for storing the results of MapReduce tasks in Hadoop. In addition, it has a specially created open-source language Pig, which can be used for some editions of MapReduce. Another instrument is Hive – database with open source, initially created for Facebook, which allows for analytical modeling in Hadoop. The most commonly used software tools that are part of Hadoop are summarized in Table 2.

Table 2.

Main software tools in Hadoop

Software Tool	Description
Apache Hive	System for data storage, based on Hadoop. Allows the user to write tasks in a language, similar to SQL, called HiveQL, which are then converted for MapReduce.
Apache Pig	Apache Pig is a platform for program creation, which works with Apache Hadoop. The language, which this platform uses, is called Pig Latin.
Apache HBase	DBMS for non-relational databases. It works with column databases - the data is organized as families of columns.
Apache Flume	Flume is software for populating Hadoop with data. The data can come from a log file, data generated by user behavior, from clicks on links, generated from various sensors, etc.
Apache Oozie	Workflow processing system, which allows one to define a series of tasks written in MapReduce, Pig and Hive and then to connect them.
Apache Ambari	Web based selection of instruments for installation, administration and observation of Apache Hadoop clusters.
Apache Avro	Data serialization system using JSON for describing data in compact binary format.
Apache Mahout	Knowledge extraction library based on machine learning algorithms.
Apache Sqoop	SQOOP is a software tool for transferring data between relational databases in Hadoop.
Apache HCatalog	Centralized service for management and sharing of metadata for Apache Hadoop.

Source: Own elaboration

The third main concept – **non-relational databases** (NoSQL – Not only Structure Query Language) have developed as a consequence of the emergence of the processes of semi-structured and unstructured data and the need for software that can process it quick enough. Some authors (Petrov and Trifonova, 2019) point out that an important aspect of NoSQL is its flexible schemes for data management, which can be easily changed, and which do not create obstacles for already stored data (Petrov and Trifonova, 2019). Examples of NoSQL databases are HBase, MongoDB, Cassandra, Accumulo, Riak, CouchDB, DynamoDB. In Fig. 5, a summary of existing NoSQL database models is given (Mazumdar, et al. 2019).

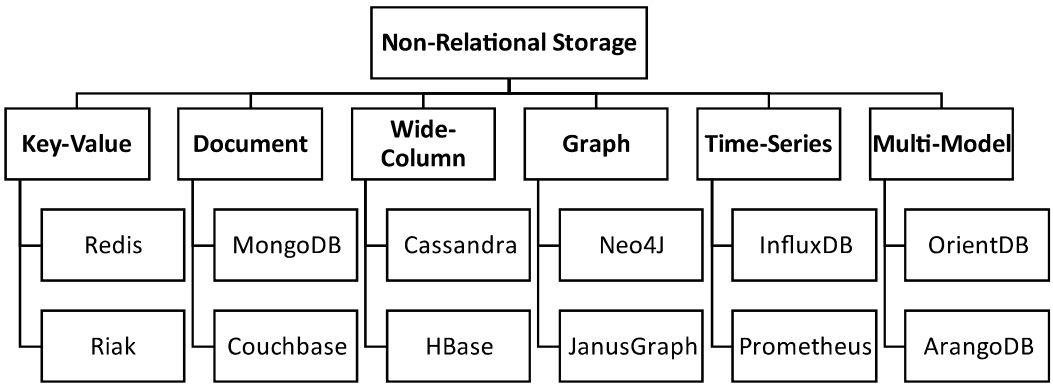

Figure 5. Classification of NoSQL
Source: (Mazumdar, et al. 2019)

NoSQL databases have characteristics that allow for scalability, work in a cloud environment and the use of parallel processing. Their other defining characteristic is that they can be used to work with real-time information flows.

Big data has also led to a change in the concepts used for data storage. Traditional repositories, based on Data Warehouse (DW) have been improved with the goal of being able to work with new types of data. The data lake (DL) concept has also appeared, which allows for the relatively inexpensive storage of different types of data, and subsequently, the application of differentiated methods of analysis.

Large volumes of fast incoming data have also led to a change in the processes of retrieval, conversion and loading of data into repositories. The traditional ETL (Extract, Transform, Load) process is when a conversion

happens. Before the data is loaded it is converted into ELT (Extract, Load, Transform) data first and then transferred to the repository where it is transformed.

In conclusion, we can point out that there are many various data sources for logistics. Among the data sources are not only complex information systems, but also real-time data from different measuring instruments, events, radio frequency identifiers, cell networks, video surveillance devices, social networks and other internet sources. The research gives us reason to conclude that in these areas there is a tendency toward an ever-increasing volume of stored data, as well as the use of data from various new sources. This leads to the need for companies to seek out and implement innovative models for the storage and processing of this data so that they can then apply appropriate modern forms of business analysis.

4. Conceptual framework for the management and processing of logistical data

The researched literature sources indicate that there are some methods, which have been developed for data processing in the logistics industry. Solutions for collecting logistics data, the completion of analysis in real-time and forecasting have been offered by a group of scientists (AlShaer, et al., 2019). The framework is called IBRIDIA and it represents a further development of their approach called ProLoD. IBRIDIA is a hybrid solution, which can be used for processing logistics data in real-time and in batch style. It allows one to collect logistics data in real-time from multiple heterogeneous sensors, social media and business processes, as well as effective data processing in real-time or in batch style and modelling and analysis of data for forecasting delays. The suggested method of IBRIDIA includes: preparation of data – collection, filtration, cleaning, integration; processing of data packets; real-time data processing; data storage.

Other authors have proposed an architecture for the platform, which is called SWeTI and they have determined that it is suitable for application in the organizations of the Industry 4.0 concept (Patel, et al., 2018). Their solution has 5 layers:

- device layer, which includes all the machines working according to an IoT standard;
- edge layer, which is based upon a network protocol for data volume and techniques for data filtration and cleaning for data refinement;
- cyber layer, which represents a center for data distribution. Stores data from different distributed sources and prepares it for processing;
- data analytic layer, distributed data lake, which facilitates analytical activities by using AI algorithms;
- application layer, presents the knowledge so that experts can make the right decisions.

In essence, the presented model is good, but we think that there is a bit of discrepancy between the name of layers and their actual function.

Another suggested solution for processing logistics data is the model of Haße and others (2019), which is a “digital twin” type. Digital twin is defined as a continuously changing digital profile, which contains historical data and new data for a given physical object or process. The main goal of this model is to optimize the effectiveness of a business, based on a large quantity of accumulated data obtained from measurements of a number of objects in the real world. The analysis of the accumulated data allows one to obtain accurate information about the operation of the system, which also provides the basis for conclusions on the necessity to make changes to the product, which is being manufactured, as well as to the manufacturing process (Parrot & Warshaw, 2020).

The model of Haße et al. applies the Lambda architecture in the logistics industry, which ensures a scalable and powerful infrastructure for collecting, processing and visualizing data from IoT. It has four layers: data acquisition; data processing with Lambda architecture; data visualization; semantic layer for digital twins in logistics. The proposed solution is flexibly scalable. The architecture operation can be adapted depending on the type and number of sensors, complexity of analysis models and the presence of computing resources in the logistics company.

The researchers of Oracle point out that with the implementation of management systems of big data in logistics, businesses will be able to achieve greater benefits and become more flexible (Oracle, 2015). They have concluded that the kind of data used, the way in which it was extracted and structured, as well as the frequency of updates and quality of the data, are crucial in deciding which is the right technology. It must be determined whether it will be used for processing in real time or in batch mode. The frequency of processing, which needs to be done based upon the availability of data, is also important. The proposed conceptual framework permits one to operate with structured, semi-structured, and unstructured data in logistics.

In another study (Lv, et al., 2020) that focused on the design of logistics parks for steel, they suggested a data-based approach aimed at improving the operational efficiency of these types of logistics structures. They

identified the frequency and correlation between the different products and used analyses to facilitate the product distribution of steel logistics parks. The data analysis module includes three steps: data pre-processing; model building and setting of model parameters.

The aim of the module is to use the obtained results to assist the operations management in order to clarify the characteristics of incoming orders, monitor the work in real-time and achieve an optimal distribution and deployment of steel products in the logistics parks.

Another solution proposed by a team of scientists (Jiang et al., 2020), which can be utilized for data processing and monitoring, includes an architecture consisting of: perception layer, network layer, big data layer and application layer. In the first layer, one finds hardware such as mobile devices, RFID, sensors, all of which collect appropriate data for the logistics process. The network layer is based on urban networks, networks for mobile communications. The data layer uses basic technologies for virtualization and distributed data storage. The application layer is divided into modules for: client relations management, warehouse management, orders and distribution management.

The data management model for logistics systems in e-commerce are proposed by another group of scientists (Zhao et al., 2020). They researched how to improve efficiency of logistics distribution in e-commerce through big data analysis. The idea of constructing a logistics model for distribution management in e-commerce includes these stages: data acquisition; data extraction and analysis; derivation of forecasts, optimization models, etc. In relation to logistics processes in e-commerce, Zampou et al. (2018) proposed a solution. In their model, they focus on security and include special connectors, which allow a higher level of security through combining various mechanisms for secure data exchange that is only between certified and connected partners. They use platforms for storage and computing of big data, which consist of a collection of network servers with resources for storage and computing working in a cluster.

This study of the existing logistics-specific solutions for processing big data gives us grounds to define the following basic principles for the construction of a modern management architecture and big data processing in this field:

1. **Adaptability** and **flexibility**. Ability to adjust to quickly changing conditions. Allowing the quick redirection of data flow in accordance with the business' goals. Every object that uses data to be utilized several times and, in this way, to ensure a constant stream of high quality, relevant data for business. The ability to maintain various types of business users, operations and the processing of different types of data.
2. **Scalability**. Variable workloads in organizations require an elastic architecture, which adapts to the changing requirements for data processing upon demand. This type of architecture allows for work to be done even when there is a change in the capacity, i.e, the use of additional applications and analytical platforms.
3. **Automation**. The processes of absorption, preparation, processing of data sets to be implemented by mechanisms that operate with minimal human intervention. So that anomalies can be discovered in real time and then to display warnings to the instituted operational control panels.
4. **Intelligent**. Necessity of the data architecture to use intelligent resources for training, settings, and data administration. Based upon algorithms for machine training, which prepares the data, identifies errors in quantity and suggests actions.
5. **Managed**. Developed by different cooperating specialists. The process of extraction, storage, and data processing is the responsibility of IT specialists as well as business analysts, who know the business context of the operations. To clearly define the different types of users and their rights to access the data.
6. **Personalized**. Compliance with the organization and its business needs. If it is a small business for example, it could work with intelligent tools in an integrated management environment. If it is a big company, it should use a system for parallel processing.
7. **Secure**. Protected from unauthorized access and compliant with the general regulation for data protection adopted by the European Union (Europa.eu, 2016).
8. **Stable**. Any data architecture should be reliable and allow recovery upon need.

On the basis of all the presented and researched methods for data processing and analysis in logistics, as well as, the principles above, we would like to offer a conceptual framework for the management and processing of data in the logistics industry (Fig. 6).

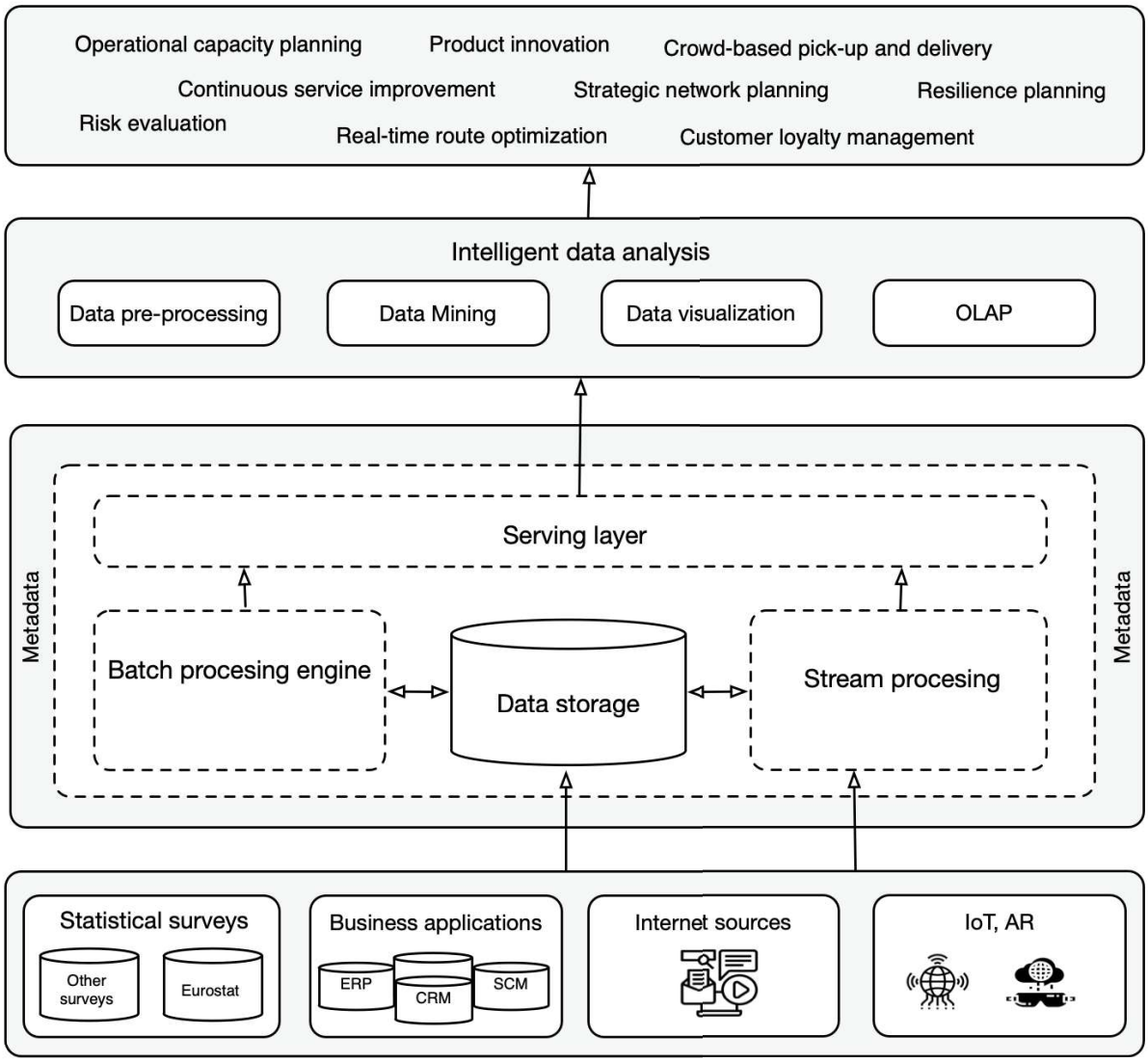

Figure 4. Conceptual framework for the management and processing of data in logistics
Source: Own elaboration

Big data processing in logistics refers to the overall process of: selection and extraction of data; storage and management of data; and the application of resources for intelligent business analyses in order to identify models, predictions, forecasts, discovery of tendencies, etc.

We have identified the following as main data sources:

- surveys;
- user-inputted data from software systems, as well as data generated from work on ERP, CRM, SCM and other systems;
- IoT devices with internet access and sensors and Augmented Reality (AR) technology;
- internet sources – server log files, webpages, social media, etc.

The data storage process should be implemented though data repositories with flexible architecture, which allows for the organization and processing of structured, semi-structured and unstructured data.

The implementation of the suggested concept can be made on the basis of various software tools for data warehouse management and business analysis. We would suggest that the implementation can be accomplished with the collection of open-source software Apache Handoop. Open-source software solutions are constantly evolving, and they allow for an easier adaptation for a specific business and the ability to make improvements. The Hadoop platform is designed for the organization of distributed processing of large volumes of data using

the map/reduce concept and for that reason, we believe that it is appropriate for the implementation of the proposed conceptual framework.

It is advisable to use the Apache Sqoop tool to transfer the data coming from the relational databases of the logistics business application in the Hadoop Distributed File System. One of its advantages is that it is easy to use and it can work with different relational database systems. Also, it was designed based on the module principle and that allows for specialized additions to be included and for optimized transfers for certain DBMSs.

Batch data processing is based on HDFS, Map-Reduce, Oozie, Pig, Hiv and Impala. HDFS distributed file system and MapReduce are the main components of Hadoop, providing parallel processing of clusters. Oozie is suitable for task planning and workflow. Pig is a platform for creating and executing tasks, it can be extended with the help of user-defined functions written in Java, Python, JavaScript and Ruby. Apache Hive is a database management system that supports SQL-based query language. Apache Impala is used for queries of data stored in the HDFS and Apache HBase.

Different sensor measurements, website and application clicks, and completed financial transactions thankfully provide an almost constant flow of data in logistics. The Apache Kafka software platform is used to receive and process real-time streaming data. It was initially designed to work with LinkedIn data, but later it transformed into a highly effective system for processing many types of streaming data. It can be used with data from websites, social networks and monitoring systems. Apache Flume is most often used for the collection and summary of streaming data, such as log files, but it must be noted that Apache Flume is not limited to working with log files only. It is used for transporting large quantities of data for events, network traffic, social media data, email messages, etc. Apache Kafka and Flume are both reliable and provide guarantees against the loss of data.

Apache Spark open-source platform can be used to achieve real-time processing and the necessary performance for distributed processing of semi-structured and unstructured streaming data. Its main advantage is speed, it can complete tasks much faster than MapReduce. Apache Spark Streaming can work with real-time streams by diverting input data streams into packets, which are processed by the Spark Engine and then it generates a final stream. It integrates well with Spark SQL and Spark MLib. It is suitable for processing data from IoT devices.

Streaming data can also be accomplished using Apache Flink – distribution mechanism for streaming data, which provides libraries for batch and streaming processing, machine learning and graphic processing. During the processing of the stream, each stream has input from one or more sources and sends the data to output streams (file or database). The data is transformed while in the stream so that the data will be suitable for real-time analyses.

In order to service the data warehouse requests, it is best to use these technologies: Cassandra, Impala, Pig Table Export, SqoopExport, Hive Tables, Impala Tables.

Intelligent analyses of big data in logistics can be performed using a large variety of data mining technologies. For this purpose, the R and Python scripts were specially designed or one can use other data mining software tools that are suitable for working with big data in the logistics industry. Some of the appropriate software tools, which the research has shown to be widely used (Landset et al., 2015) are:

- MLlib – Spark library for machine learning. It uses algorithms for data pre-processing and transformation, classification, regression, and clustering (spark.apache.org, 2021);
- Apache Mahout – framework for creating algorithms focused on linear algebra for machine learning (mahout.apache.org, 2021);
- SAMOA (Scalable Advanced Massive Online Analysis) – open-source platform for data mining from large data streams. Offering a collection of distributed streaming algorithms for the most often encountered data mining and machine learning tasks such as classification, clusterization and regression;
- H2O – open-source platform for machine learning with memory with linear scalability. It works with Handoop, which provides the ability for data importation from various sources and it has a quick, scalable and distributed computing mechanism written in Java. It supports the most commonly used statistics and algorithms for machine learning. The platforms includes interfaces for R, Python, Scala, Java, JSON and CoffeeScript/JavaScript and built-in Flow web interface;
- RapidMiner Radoop – machine training for Hadoop and Spark. It provides an easy-to-use graphic interface for analysis of data in the Hadoop cluster with the Hive working server.

The proposed conceptual framework and its implementation fully comply with the above defined principles for constructing a modern management architecture and the processing of big data in logistics.

5. Conclusion

The collection and processing of data from diverse and heterogeneous sources in the logistics industry creates problems for conventional logistics information systems. Existing software solutions are not able to process large volumes of data from sensor devices or social media in real-time. Our research proposes a conceptual framework for the processing of specific structured, semi-structured and unstructured data in logistics. The framework would provide logistics companies with many advantages, such as forecasting events, route optimization in real-time, prevention of unexpected delivery delays, implementation of innovative solutions and an overall adaptation to digitalization processes.

Literature

Alessia Napoleone, A., Macchi, M., Pozzetti, A. (2020). A review on the characteristics of cyber-physical systems for the future smart factories. Journal of Manufacturing Systems, 54, pp. 305–335.

AlShaer, M., Taher, Y., Haque, R. & Hacid, M., 2019. IBRIDIA: A hybrid solution for processing big logistics data. Future Generation Computer Systems, Том 97, pp. 792-804.

Bahga, A. and Madisetti, V. (2019). Big Data Analytics: A Hands-On Approach. Published by Arshdeep Bahga & Vijay Madisetti.

Chaopaisarn, P. and Woschank, M. (2020). Maturity Model Assessment of SMART Logistics for SMEs. Chiang Mai University Journal of Natural Sciences. 20(2), pp. 1-8.

Cio.com, 2019. AI Unleashes the Power of Unstructured Data. [Online] Available from: https://www.cio.com/article/3406806/ai-unleashes-the-power-of-unstructured-data.html. [Accessed 06/02/2021].

Dimitrova, V., Semova, M., 2018. Blockchain and Smart Contracts in Retailing. Commerce 4.0 – Science, Practice and Education: International Scientific Conference, pp. 126-136.

Domo.com, 2021, Data Never Sleeps 8.0, [Online] Available from: https://www.domo.com/learn/data-never-sleeps-8 [Accessed 07/02/2021].

Emilova, P., 2016. Economic aspects of the use of cloud services. (in Bulgarian). Anniversary scientific conference "Challenges to information technology in the context of Horizon, Svishtov: Tsenov Academic Publishing House, pp. 126-132.

Europa.eu, 2016. Data protection in the EU. [Online] Available from: https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_bg [Accessed 07/02/2021].

Eurostat, 2021. Cloud computing – statistics on the use by enterprises. [Online] Available from: https://ec.europa.eu/eurostat/statistics-explained/index.php/Cloud_computing_statistics_on_the_use_by_enterprises. [Accessed 04/02/2021].

Filipova, N., 2018. Blockchain – an opportunity for developing new business models. Business management, SA „D. A Tsenov“ Svishtov, 2, pp. 76-92.

Firican, G., 2017. The 10 vs of big data, upside where data means business. [Online] Available from: https://upside.tdwi.org/Articles/2017/ 02/08/10-Vs-of-Big-Data.aspx?Page=1. [Accessed 06/02/2021].

fortunebusinessinsights.com 2020. Market Research report. [Online] Available from: https://www.fortunebusinessinsights.com/industry-reports/big-data-technology-market-100144 [Accessed 07/02/2021].

Gergova, I. et al. 2017. Possibilities of GIS for optimization of localization processes in tourism. Management and sustainable development, 4 (65), [Online] Available from: http://oldweb.ltu.bg/jmsd/files/articles/65/65-05_I_Gergova_Minov_Momchev_Kovacheva_Ivanov.pdf

Ghobakhloo, M., 2020. Industry 4.0, digitization, and opportunities for sustainability. Journal of Cleaner Production 252, 119869, [Accessed 07/02/2021].

Haße, H. et al., 2019. Digital twin for real-time data processing in logistics. Chapters from the Proceedings of the Hamburg International Conference of Logistics (HICL), Institut für Logistik und Unternehmensführung, Technische Universität Hamburg, Том 27, pp. 4-28.

Kirov, S., 2020. Blockchain applications in the insurance industry. (in Bulgarian). Jubilee International Scientific Conference “Economic science, education and the real economy: development and interactions in the digital age” T. 1, Varna: Publishing house "Science and Economy", pp. 186-200.

Kondratiev, V., 2019. The fourth industrial revolution and globalization. (in Bulgarian) [Online] Available at: https://geopolitica.eu/2018/174-broy-6-2018/2941-chetvartata-industrialna-revolyutsiya-i-globalizatsiyata. [Accessed 3.02.2021].

Kotu, V. and Deshpande, B. (2019). Data Science. Concepts and Practice Second Edition. Morgan Kaufmann Publishers an imprint of Elsevier.

Lazarova, V., 2019. Digitalization and Digital Transformation in Accounting. Ikonomiceski i Sotsialni

Alternativi, Sofia: University of National and World Economy, issue 2, pp. 97-106.

Lv, Y., Xiang, S., Zhu, T. & Zhang, S., 2020. Data-Driven Design and Optimization for Smart Logistics Parks: Towards the Sustainable Development of the Steel Industry. MDPI, Open Access Journal, 12(17), pp. 1-13.

Lydon, B., 2014. Industry 4.0 - Only One-Tenth of Germany's High-Tech Strategy. [Online] Available at: https://www.automation.com/en-us/articles/2014-1/industry-40-only-one-tenth-of-germanys-high-tech-s. [Accessed 3.02.2021].

Mazumdar, S., et al., 2019. A survey on data storage and placement methodologies for Cloud-Big Data ecosystem. Journal of Big Data, Vol. 6, pp. 1-37.

mi.government.bg, 2017. Concept for digital transformation of the Bulgarian industry (Industry 4.0). (in Bulgarian) [Online] Available at: https://mi.government.bg/bg/themes/koncepciya-za-cifrova-transformaciya-na-balgarskata-industriya-industriya-4-0-1862-468.html [Accessed 3.02.2021].

Oracle, 2015. Improving Logistics & Transportation Performance with Big Data. [Online] Available at: https://wheels.report/Resources/ Whitepapers/9b500313-465f-48cb-90d1-719404791108_C.pdf [Accessed 04/03/2021].

Paksoy, T. Koçhan, Ç. Ali, S., 2021. Logistics 4.0. Digital Transformation of Supply Chain Management. CRC Press Taylor & Francis Group.

Parrott, A. & Warshaw, L., 2020. Industry 4.0 and the digital twin technology. [Online] Available at: https://www2.deloitte.com/us/en/insights/focus/industry-4-0/digital-twin-technology-smart-factory.html [Accessed 04/03/2021].

Patel, P., Sheth, A. & Ali, M., 2018. From Raw Data to Smart Manufacturing: AI and Semantic Web of Things for Industry 4.0. Intelligent Systems, IEEE, 33(4), pp. 79-86.

Petrov, Ph. and Trifonova, Ts., 2019. Non-relational databases – a practical guide. Sofia: University Publishing House "St. Kliment Ohridski ”.

Proctor, M. and Wilkins, J., 2019. 4.0 Sight—digital industry around the world. Technical report, EU Automation. [Online] Available at: http://www.4sightbook. com/. Accessed Feb 2020 [Accessed 04/02/2021].

pwc.de, 2017. Digital Factories 2020: Shaping the future of manufacturing. [Online] Available from: https://www.pwc.de/de/digitale-transformation/digital-factories-2020-shaping-the-future-of-manufacturing.pdf [Accessed 03/02/2021].

Rojko, A., 2017. Industry 4.0 concept: background and overview. International Journal of Interactive Mobile Technologies. (iJIM), 11(5): pp. 77-90.

Satatista.com, 2021. Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2024. [Online] Available at: https://www.statista.com/statistics/871513/worldwide-data-created/, [Accessed 3.02.2021].

Shafer. T., 2017. The 42 V’s of Big Data and Data Science. [Online] Available from: https://www.kdnuggets.com/2017/04/42-vs-big-data-data-science.html [Accessed 04/02/2021].

Sharda, R. Delen, D. Turban, E., 2020. Analytics, Data Science, & Artificial Intelligence Systems for Decision Support. Pearson Education.

Sharda, R., Delen, D., Turban, E., 2020. Analytics, data science, & artificial intelligence systems for decision support. Pearson.

Statista.com, 2021. Number of Internet of Things (IoT) connected devices worldwide in 2018, 2015 and 2030. [Online] Available from: https://www.statista.com/statistics/802690/worldwide-connected-devices-by-access-technology/ [Accessed 04/02/2021].

Storeya, V. and Song, I., 2017. Big data technologies and Management: What conceptual modeling can do. Data & Knowledge Engineering, 108, pp. 50-67.

Stoyanov, S. and Popchev, I., 2017. Internet of Things. (in Bulgarian) [Online] Available at: https://www.researchgate.net/profile/Ivan_Popchev/publication/322953157_Internet_na_nesata/links/5a79909445851541ce5ce863/Internet-na-nesata.pdf [Accessed 3.02.2021].

Stoyanova, M., Vasilev, J., Cristescu, M. 2021. Big Data in Property Management. Applications of Mathematics in Engineering and Economics: Proceedings of the 46th Conference on Applications of Mathematics in Engineering and Economics (AMEE '20), AIP Conference Proceedings 2333, 070001.

Turchi, P., 2018. The Digital Transformation Pyramid: A Business-driven Approach for Corporate Initiatives. [Online] Available from: https://www.linkedin.com/pulse/digital-transformation-pyramid-business-driven-approach-turchi [Accessed 03/02/2021].

Vasilev, J., 2017. E-logistics in the context of globalization. (in Bulgarian) Varna: Publishing house "Science and Economy".

Zampou, E. et al., 2018. Big data analytics in e-commerce logistics: Findings from a systematic review and a

case study. Proceedings of 7th Transport Research Arena TRA 2018, Vienna, Austria, [Online] Available at: https://zenodo.org/record/1491581#.X6fU5y8RoXo. [Accessed 8.11.2020].

Zhang, D. et al., 2021. A survey on attack detection, estimation and control of industrial cyber-physical systems, ISA Transactions. [Online] Available from: https://www.sciencedirect.com/science/article/pii/S001905782100046X [Accessed 04/02/2021].

Zhao, Y. et al., 2020. Innovation Mode and Optimization Strategy of B2C E-Commerce Logistics Distribution under Big Data. Sustainability, 12, 3381, pp. 1-13.