JOURNAL OF SECURITY AND SUSTAINABILITY ISSUES
ISSN 2029-7017 print/ISSN 2029-7025 online
2022 Volume 12
http://doi.org/10.47459/jssi.2022.12.12

ASSESSMENT OF IT SYSTEMS ARCHITECTURE IN THE CONTEXT OF BIG DATA PROCESSING FOR SMART CITIES DEVELOPMENT

Hubert Szczepaniuk[1], Edyta Karolina Szczepaniuk[2]

[1] Warsaw University of Life Sciences, Nowoursynowska 166 ST., 02-787 Warsaw, Poland
[2] Polish Air Force University, Dywizjonu 303 no. 35 ST., 08-521 Dęblin, Poland

E-mails:[1] hubert_szczepaniuk@sggw.edu.pl; [2]e.szczepaniuk@law.mil.pl

Received 15 May 2022; accepted 9 September 2022; published 30 September 2022

Abstract. Effective use of Big Data can significantly support the development of smart cities and the new digital economy. The aim of the article is a multi-criteria evaluation of IT systems in terms of Big Data processing, taking into account the support for the development of smart cities. The article includes theoretical and empirical research. The adopted criteria for assessing the architecture of IT systems relate to barriers to the implementation of the digital economy in smart cities and the guidelines of international data strategies. The evaluation covered, among other things, cybersecurity and the effectiveness of organizing, storing, and producing new information. The research results allowed us to identify the key factors of Big Data processing efficiency. Based on the research results, an effective model of Big Data processing in organizations was developed. In particular, various data models were analyzed as one of the main elements of software architecture of information systems. The research also focused on data processing techniques such as data warehousing, machine learning, and distributed computing. The efficiency factors of IT systems identified in the research reduce barriers to developing global data strategies and smart cities.

Keywords: Big Data; IT Systems; Cybersecurity; Smart Cities; Software Engineering; Data Models; Mathematical Model; NoSQL;

Reference to this paper should be made as follows: Szczepaniuk, H., Szczepaniuk, E.K. 2022. Assessment of it systems architecture in the context of big data processing for smart cities development. *Journal of Security and Sustainability Issues,* 12(1), 139-158. http://doi.org/10.47459/jssi.2022.12.12

JEL Classifications: C61, C88, D83, L86

Additional disciplines: Computer Science; Security Sciences

1. Introduction

One of the fundamental postulates of sustainable development is promoting the concept of smart cities and the new digital economy. The development of smart cities involves using various information and communication technologies (ICT) to increase the quality and efficiency of the available urban infrastructure (e.g. Korneć, 2020; Zhao et al., 2020; Koman et al., 2022). From a technical point of view, the functioning of smart cities requires the use of data collected through a variety of sensors to automate services, improve efficiency, and reduce costs and environmental damage (Ullah et al., 2021).

The broadly understood development of digital services implemented in smart cities means the need to efficiently process data collected using various sensors and Internet of Things (IoT) devices. The data collected and processed in the various functional sectors of smart cities have a heterogeneous structure that changes over time and often has to be processed in real-time. Therefore, IT systems supporting digital services in smart cities operate in Big Data conditions.

Key organizations supporting socio-economic development emphasize that Big Data will be one of the fundamental pillars of sustainable development in the future. The European Union implements Big Data-oriented strategies in shaping the digital future of Europe, which translates into increased productivity of all sectors of the economy and supports sustainable development (*Big data*, website state as of May 6, 2022; *A European Strategy for data*, website state as of May 6, 2022).

In the source literature, Big Data's definition is often based on indicating the key attributes of such datasets. The classic definition of Big Data is based on the 3V model, which includes such attributes as volume, velocity, and variety (Laney, 2001). The first attribute indicates large volume datasets, challenging to store in classic database systems. The velocity attribute refers to requirements within the aspect of time necessary for processing data. Big Data often flow in streams, and their processing is carried out in time close to real. The variety attribute emphasizes the occurrence of unstructured data with an incoherent structure. The 3V model was expanded to the 5V model, in which the veracity and value attributes were added (Lomotey & Deters, 2014). The veracity attribute considers requirements in the aspect of data integrity and authenticity. The value attribute establishes the requirement of defining criteria for selecting data worth being analyzed.

It is, therefore, crucial to find a solution to the problem of creating an IT system that will efficiently process Big Data, considering the implications for the development of smart cities and international data strategies. Cox and Ellsworth (1997) presented a layered visualization and data management model, including visualization algorithms, data models, data management, and distributed data storage. The expressed concept constitutes a technical approach enabling analysis of the key components of the architecture of IT systems within the context of Big Data processing. The IT perspective can be extended to a smart city development perspective and international data strategies. Effective use of Big Data can promote the development of smart cities by optimizing and automating processes and supporting decision-making.

2. Related works

Big Data is an interdisciplinary concept and can be considered on many levels. For the purposes of this paper, a literature review was made on supporting selected functional sectors of smart cities through the use of Big Data. Knowledge discovered through Big Data processing is the key to improving quality, implementing innovation, better personalization, and making city services smarter. On the other hand, many Big Data projects in organizations operating in smart cities fail (see e.g., Barham & Daim, 2020). Also, for this reason, the recognition of the factors of the effectiveness of Big Data processing by the organization's IT systems is crucial. The literature review related to the topic was carried out, taking into account the key functional areas of smart cities (Table 1).

Table 1. The benefits of using Big Data in selected functional areas of smart cities

Selected functional sectors of smart cities	Application areas of Big Data	Source
Security	- improving the privacy and security of data management in smart cities - increasing citizens' rights to access data	Chen et al. (2021)
Environmental protection	- municipal waste management - support for the green supply chain - improving regional environmental management - promoting the ecological transformation of hospitals - assessment of the perception of environmental pollution by multiple stakeholders	Limba et al. (2020) Liu et al. (2022) Benzidia et al. (2021)
Healthcare	- better diagnosis and prevention of diseases - improving the quality of medical services - supporting the semantic interoperability of medical data - promoting the concept of intelligent hospitals - the use of IoT ecosystems to support medical processes	Pustochin et al. (2021) Ullah et al. (2017) Karatas et al. (2022)

Public transport	- discovering users' hidden travel patterns - dynamic optimization of vehicle resources - promotion of sustainable transport - supporting the concept of the sharing economy - supporting the low-carbon smart green tourism	Bi & Ye (2021) Ma et al. (2021)
Energetics	- intelligent thermostats for building heating control - energy conservation for IoT ecosystems - supporting the energy efficiency of smart cities	Esrafilian-Najafabadi, & Haghighat (2021) Khan et al. (2017) Anthopoulos & Kazantzi (2022)

Source: Own work based on a literature review.

The effective use of Big Data can significantly contribute to increasing the level of privacy and security in sustainable cities. Chen et al. (2021) proposed a holistic approach to intelligent Big Data modeling to improve the privacy and security of data management in smart cities. The authors reasonably point out that Smart City systems require different data protection methods against cybercrime and malicious attacks on data management servers. From the point of view of this article, it is essential that the cited studies indicate that smart city software generates huge amounts of data in various formats coming from various domain areas, such as traffic, electricity, education, health care, and development. In this article, the factors that affect data processing efficiency in various formats will be analyzed further.

Environmental protection in sustainable cities can be significantly improved through the effective use of Big Data. Limba et al. (2020) analyzed Big Data applications in the municipal waste management sectors and cryptocurrencies as elements of Industry 4.0. The research findings of the cited studies indicate the possibility of effective application of Big Data in both sectors. However, the authors identified both positive and negative elements and barriers to implementing Big Data (see e.g., Limba et al. 2020). In turn, Liu et al. (2022) conducted multi-stakeholder perception research on environmental pollution based on Big Data. In the cited studies, the authors used Big Data to assess the perception of the environment by various stakeholders (society, government, media, companies, and scientists). Thanks to the conducted analyzes, it is possible to improve the regional environmental management to develop sustainable cities.

A key area of smart cities is healthcare. Big Data analyzes in healthcare can significantly improve medical services, earlier disease detection, and better prevention. Pustochin et al. (2021) proposed to optimize the deep learning method and Big Data analysis in healthcare for 5G mobile networks. In particular, the authors developed a new feature selection concept based on Big Data analysis and the Deep Belief Network disease diagnostic model. In the context of the research carried out in this article, it is crucial that the cited authors emphasize the significant heterogeneity of data in the healthcare sector, which occurs due to the integration of various biomedical data sources. In particular, quantitative data such as sensor data, gene matrices, laboratory samples, and qualitative data, e.g., text, were highlighted. Data heterogeneity as one of the key features of Big Data will be further explored in this article.

Ullah et al. (2017) presented a semantic interoperability model for Big Data based on IoT applications in healthcare. Research has shown that semantic data analysis makes it possible to identify hidden patterns in Big Data sets. Thanks to this, it is possible to implement an intelligent cloud that supports healthcare in terms of treatment strategies and counteracting the symptoms of various diseases.

Karatas et al. (2022) indicated how Big Data changed the nature of healthcare in the era of Industry 4.0. The authors highlighted areas of IoT systems in healthcare that enabled the implementation of complete ecosystems consisting of intelligent sensors and cloud services. Based on the research carried out in the cited article, it was concluded that Industry 4.0 is an integral part of future solutions in smart factories and intelligent hospitals. Thanks to Big Data analysis, software, technologies, and processes can generate better results while minimizing time and costs.

Big Data analysis can also significantly support the development of public transport in sustainable cities. Bi &

Ye (2021) developed a method based on Big Data analysis to discover the hidden patterns of user travel. By discovering user behavior, it is possible to optimize vehicle resources dynamically. As a result, research findings promote sustainable transport and support the sharing economy concept to develop an environmentally friendly community. Various Big Data analysis methods were used in the cited studies, including pattern analysis, cluster analysis using k-means clustering based on POI data, and the BoW (Bag-of-Words) algorithm. In this article, various Big Data methods and techniques will be subjected to a multi-criteria analysis.

Ma et al. (2021) used Big Data to support low-carbon smart green tourism. The authors analyzed the impact of big data marketing techniques on the durability of low-emission supply chains in tourism. Research results show that Big Data analysis enables the creation of intelligent and personalized recommendations for low-carbon tourism. In terms of environmental protection, the research results obtained allow low-emission tourism to become more intelligent and promote the concept of environmental protection among tourists. The obtained research results contribute to making optimal decisions in green tourism, taking into account the purchasing behavior of tourists regarding low-emission services.

Big Data analytics is also used in the energy sector of smart cities. Khan et al. (2017) proposed an energy-saving IoT architecture that can be implemented in smart homes and cities. The authors indicated that planning is one of the ways to minimize IoT ecosystems' energy consumption. The solution presented in the cited article is based, among other things, on the analysis of data obtained from sensors monitoring the operation of IoT devices so that they can be intelligently turned off in order to save energy. The research used the Big Data analysis method based on the MapReduce paradigm and the Hadoop distributed file system. The indicated techniques of Big Data analysis will be examined in this article in the context of effectiveness.

Anthopoulos & Kazantzi (2022) defined the taxonomy of energy efficiency and analyzed the models of its evaluation in cities in terms of artificial intelligence and big data. The authors emphasized that energy efficiency is a key issue for the sustainability of today's and tomorrow's cities. It has been reasonably argued that new technologies implemented in smart cities such as blockchain, electric vehicles, autonomous vehicles, smart buildings, artificial intelligence, and Big Data challenge identified urban energy efficiency because they require enormous amounts of power. The results indicate that more research is needed to reduce AI and Big Data energy consumption to improve the profits generated by these technologies. In the field of energy-saving, there are practical implementations of the use of artificial algorithms that bring real benefits. Esrafilian & Haghighat (2021) developed a deep learning HVAC control method for online pre-conditioning time estimation in buildings.

Finally, Big Data can contribute to the improvement of environmental protection. Benzidia et al. (2021) analyzed the impact of Big Data and artificial intelligence on the integration of green supply chain processes and the environmental performance of hospitals. The authors have shown, among other things, that decisions based on Big Data and artificial intelligence have a positive impact on green supply chains and the integration of environmental processes, which makes it possible to improve environmental performance. The research shows that Big Data and artificial intelligence allow for the adoption of an effective decision-making model to select suppliers based on environmental criteria and promote the ecological transformation of hospitals.

3. Research methodology

The aim of the article is a multi-criteria assessment of selected components of IT systems architecture in terms of Big Data processing, taking into account international data strategies and the development of smart cities. For this purpose, the key factors determining the effectiveness of Big Data processing in IT systems of modern organizations were identified.

The following specific objectives were distinguished in the research:

1) Development of meta-model requirements for Big Data processing in relation to the digital economy and smart cities requirements.

2) Assessment of IT systems architecture in Big Data processing based on the results of empirical research.

3) Identification of the key factors of Big Data processing efficiency.
4) Defining an effective Big Data processing model.

The article uses research methods referring to software engineering, information systems modeling, mathematical modeling, empirical diagnostic survey, and statistical analysis.

Theoretical considerations concern the identification of scientifically justified requirements for selected components of IT systems architecture in the field of effective Big Data processing. In particular, data models were analyzed as one of the main components of software architecture of IT systems. The research included both the traditional relational model approach as well as NoSQL data models. The research also focused on data processing techniques such as data warehousing, machine learning, and distributed computing. Another key element of the research was the cybersecurity assessment of Big Data processing. Moreover, the adopted criteria for assessing IT systems in organizations relate to information management strategies and challenges in implementing the digital economy.

The theoretical analyzes carried out made it possible to determine the dependent and independent variables for empirical research. The empirical research aimed to identify the key factors of Big Data processing efficiency in organizations. Finally, the holistic approach to the research topic allowed to define the model of effective Big Data processing in organizations supporting the development of smart cities.

4. Theoretical basis of research – Big Data processing metamodel in organizations

According to the systemic approach, the information system can be seen as a multi-level structure that allows the user to process input data into output valuable information from the decision-making point of view. However, it should be pointed out that in the literature, one can find research approaches emphasizing the multidimensionality of the organization's information system. In particular, technical, social, socio-technical, and process perspectives can be distinguished (Boell & Cecez-Kecmanovic, 2015). On the technical side, an information system can be defined "as a set of interrelated components that collect (or retrieve), process, store, and distribute information to support decision making and control in an organization" (Laudon & Laudon, 2012, p. 15). On the other hand, the technical perspective mainly concerns hardware and software, including computer networks, databases, and data analysis algorithms. In the general perspective, the organization's information system realizing Big Data processing can be interpreted as a function *f* realizing the transformation of Big Data into valuable business knowledge at the operational, tactical, or strategic management level. The implementation of the function *f* in the general framework of the Big Data processing metamodel should include data models, analytical procedures and algorithms, and the organization of information processes in line with international data strategies. The framework metamodel of Big Data processing in organizations is presented in Figure 1.

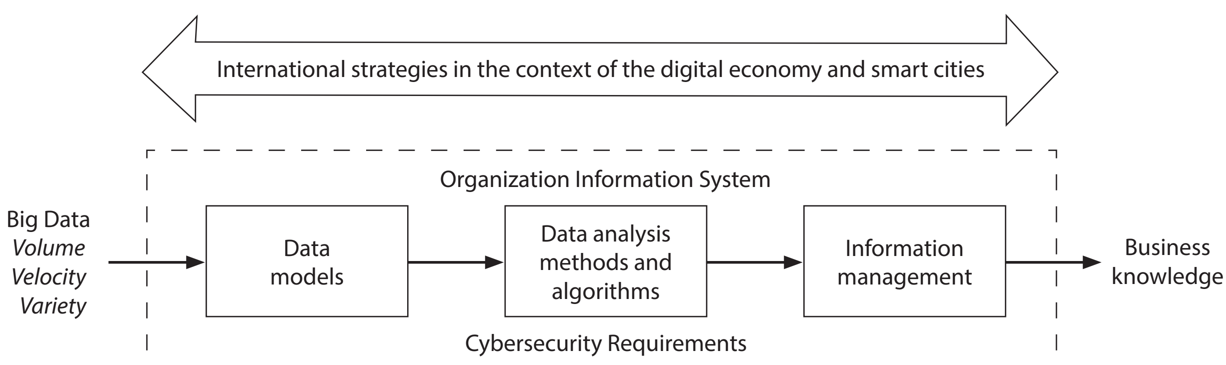

Figure 1. Big Data processing metamodel in organizations

Source: Own work.

4.1. International data strategies in the context of developing the new data economy and smart cities

International data strategies emphasize the importance of effective Big Data processing in creating social progress and economic growth. The results of the public consultation, published on the European Commission's website, show that businesses and citizens have opted for a data-driven fair economy (*Data Act: Businesses and citizens in favour of a fair data economy*, website state as of April 9, 2022). A public consultation was carried out in 2021 and focused on the importance and value of data for business and individual consumers (*Data Act: Businesses and citizens in favour of a fair data economy*, website state as of April 9, 2022). Many guiding principles for data policy have been adopted to implement the data management strategy in the EU. In particular, these principles relate to data management, interoperability, data standards, data quality, and data cybersecurity (see e.g., European Commission, 2020). Implementing the postulates of European data spaces provides greater access to common data sets that can be effectively used in the economy, and society (*A European Strategy for data,* website state as of May 6, 2022).

The UN's Sustainable Development Goals include the development of smart cities in a safe and sustainable manner (see e.g., *Goal 11: Make cities inclusive, safe, resilient and sustainable*, website state as of May 19, 2022). To achieve the assumed postulates, cities implement intelligent technologies that help to face various crises and improve the quality of services. However, the development of the smart city concept also faces many barriers. In terms of the development of data platforms in smart cities, the following barriers can be mentioned in particular:
– the problem of unifying heterogeneous data streams from various sectors of the city's functioning;
– the problem of ensuring effective data analysis to make decisions supporting the development of the city;
– the problem of interoperability of systems providing services in smart cities;
– data protection and privacy issues.

The above barriers can be minimized through appropriate Big Data processing strategies by the organization's IT systems. In particular, it is crucial to consider all stages of the IT system life cycle, starting from the early stage of software modeling and programming. It is also indicated that mutual understanding between different stakeholders (managers, users, and programmers) influences the project's success (Jenkin, Chan, & Sabherwal, 2019).

4.2. Data models

The selection of a data model at the design stage of an IT system directly determines the effectiveness of the entire system in its life cycle. A properly selected data model translates into the effectiveness of Big Data processing in terms of the relations' context with implemented business processes. Computer Science provides various concepts in data models and database management systems. In the source literature, there is a division into relational databases and No-SQL class databases (e.g., Szczepaniuk & Szczepaniuk, 2019). Relational databases include a classic formalized data model based on the mathematical basis of relationships; whereas NoSQL class databases represent a wide range of solutions other than the relational model. Both traditional relational databases and NoSQL databases are used in Big Data processing. Currently, relational database technologies are the most popular in implementing business solutions. According to the Database Management System popularity ranking, the first four places are occupied by relational databases, and the fifth is taken by a document-oriented database (DB-Engines Ranking, 2022). Three key database solutions were selected for further research:
– relational databases,
– document-oriented databases,
– graph databases.

The formal theoretical model of relational databases dates back to the 1870s, and E. F. Codd, an IBM employee, developed it. Codd (1972) based on the mathematical concept of a relationship determined on the extended Cartesian product. Based on the formal model, relational databases ensure data integrity and compliance with the ACID properties (atomicity, consistency, isolation, and durability). The high level of formalization of the data model carries limitations on processing files that meet the Big Data properties, especially regarding the speed and diversity attributes. Problems with the response time of complex SQL queries in relational databases and the lack of a flexible data model are well known.

The indicated restrictions of the relational model and the information needs of modern enterprises often implemented in the conditions of heterogenous datasets forced IT to search for alternative models that will offer greater flexibility both at the stage of system implementation and in the subsequent processing and analysis of datasets (see e.g., Szczepaniuk and Szczepaniuk, 2019). As a result of these requirements, NoSQL database technologies have emerged that can support the volume, variety, and velocity attributes of Big Data. From the point of view of Big Data, document-oriented databases and graph databases deserve special attention. Document-oriented databases store data in documents identified by keys. Depending on specific implementations, the structure of documents may take various formats, e.g., XML, JSON, or BSON. The lack of imposed data schema is perfect for any modifications of the data model when processing Big Data sets regarding the variety attribute. Document databases make it easy to build, extend and modify the data model at any system implementation stage. However, it must be noted that maintaining the coherency and integrity of a database depends on the application. Document databases also do not guarantee reliability due to a lack of compliance with the ACID approach.

Graph databases are based on the mathematical concept of the direct multi-graph using its properties in vertices and edges, which store data entities and relations, respectively. The mathematical properties of a graph enable it to be effectively and quickly searched. The time required to process SQL queries with multiple JOINs may be significantly longer than the time needed to query the same data in a graph database (Szczepaniuk, 2017). Another advantage of a graph database is the ability to implement an additional layer of semantics. Relations between graph nodes can increase the efficiency of machine learning and artificial intelligence algorithms (see e.g., Szczepaniuk and Szczepaniuk, 2019).

The characterized data models have specific features as well as possibilities and limitations related to their implementation in Big Data. The comparison of data models in Big Data processing is presented in Table 2.

Table 2. Overview of chosen data models in Big Data

Data model	Components	Opportunities and limitations in Big Data processing	
Relational	– data tables; – primary and foreign keys; – relations between tables;	Opportunities	– suitable for statistical analyses – compliance with ACID – easy storing and processing of structured data – the high-quality data source for data warehouses, distributed computing, and Machine Learning;
		Limitations	– need for relational-object mapping using object-oriented programming languages; – difficult storing of unstructured data; – long time required to download data from the database; – restrictions on embedded data types and the length of data fields;
Document	– documents in JSON, BSON, XML formats; – document collections;	Opportunities	– an additional layer of database semantics resulting from the relations in a graph; – fast data search based on mathematical graph searching algorithms; – relations within a database-oriented; – transaction support will enable integration with other databases;
		Limitations	– a large number of transactions challenging to process; – difficult to implement analytical queries regarding an entire database; – possible restrictions regarding concurrent and parallel data access operations;

Graph	– graph vertices; – graph edges;	Opportunities	– the flexibility of the data model and schema due to the lack of a strictly defined data structure; – good cooperation with object-oriented application code due to lack of need for ORM mapping; – primary and secondary indexing mechanisms to increase data search speed; – embedding data in documents allows a smaller number of queries to download large amounts of data;
		Limitations	– lack of standardized one query language; – complex integration with data warehouses due to possible data incoherency; – data redundancy possible;

Source: Own work.

The effectiveness of an organization's information system and business data analysis algorithms directly depends on the data models used in operational database systems.

4.3. Data analysis procedures and algorithms

Another metamodel element for processing Big Data in an organization involves extracting knowledge and information from data useful for management and decision-making. Three key technology classes are distinguished: data warehouses, distributed processing, and machine learning.

Data warehouse (DW) may constitute a basis for business analysis and support decision-making processes in an organization that processes large datasets. DW have proven their usefulness in Big Data processing and supporting managers in making decisions considering a whole organization (Bhatia et al., 2019). DW enables the integration, storage, and analysis of business data from various subsystems. One of the basic functionalities of DW is the ability to integrate and standardize data stored in the company based on various models (e.g., relational, document, column, chart) and data from poorly structured sources, e.g., from flat files. For this purpose, DW uses the ETL mechanism (Extraction, Transformation, and Load). DW contains clean, integrated data from various sources transformed into a uniform structure (Bhatia et al., 2019). In DW, data is modeled in a schema consistent with a multidimensional cube concept, followed by analytical processing queries (OLAP) that help decision-making (Scabora et al., 2016). DW use BI tools, data mining, and reporting services to transform input data into business information. Some researchers note that due to technological progress, classic data warehouses are less and less suitable for storing business data with the current demand of companies, especially in Big Data processing (see e.g., Panwar & Bhatnagar, 2020). Among the main reasons for this are limitations related to the ETL process, where analytical plans should be known before loading data into warehouses (see e.g., Panwar and Bhatnagar, 2020). For this reason, an approach towards concepts that synergistically combine the key assumptions of data warehouses and distributed processing is required.

It is widely recognized that the key to success in Big Data processing are efficient parallel and concurrent computation algorithms while meeting scalability and performance requirements (Maitrey & Jha, 2015). Recently, the MapReduce framework, which enables the fulfillment of the indicated requirements, is gaining popularity. A significant number of researchers point to the significant benefits of using parallel processing paradigms based on the Apache Hadoop MapReduce programming platform in processing Big Data (see e.g., Lin et al., 2020; Maitrey & Jha, 2015; Dittrich & Quiané-Ruiz, 2012). Hadoop MapReduce is a programming approach that allows the implementation of parallel processing algorithms using multiple computing units (nodes) organized in a cluster architecture.

Artificial intelligence (AI) is another class of technology gaining popularity in the field of Big Data processing, with particular emphasis on machine learning (see e.g., Krishnamoorthy et al., 2020; Stergiou et al., 2020; Zhou et al., 2017; Obermeyer & Emanuel, 2016; Qiu et al. 2016; Al-Jarrah et al., 2015; Xing et al., 2015). IT systems using machine learning are used in areas requiring Big Data processing, such as astronomy, biology, climatology, medicine, finance, and economy (Al-Jarrah et al., 2015). Machine learning is related to analyzing

ISSN 2029-7017 print/ISSN 2029-7025 online

learning processes and developing systems to improve their functioning based on experience (Stefanowski, 2009). In computer science, there are multiple classifications related to machine learning, which are based on different concepts in AI. However, among machine learning algorithms, the following, among others, are used: artificial neural networks, deep learning, regression algorithms, decision trees, Bayesian algorithms, clustering algorithms, and associative algorithms. Machine learning algorithms are capable of solving complex nonlinear problems. The quality of the input data determines the effectiveness of Machine learning algorithms. Machine learning algorithms can also be powered by Big Data sets, which makes it possible to use the potential of large, unstructured data sets. Machine learning algorithms can use Big Data to learn, predict or automate processes in an organization.

Some researchers point out that, despite the undoubted potential of Machine learning, most traditional techniques are not efficient or scalable enough to effectively process data that fulfills all Big Data attributes (see e.g., Qiu et al., 2016). This implies the need to conduct research on the adjustment of Machine learning to process large data sets, especially within 5V attributes. The characterized technologies have specific features as well as opportunities and limitations related to their implementation. The comparison of the analyzed technologies is presented in Table 3.

Table 3. Overview of Big Data processing technologies

Technologies	Components	Opportunities and limitations in Big Data processing	
Data Warehouse	– data sources – data integration (ETL) – analytical cube – reporting services, BI, data mining	Opportunities	– integration of data from various sources, database management systems, and flat files into a multidimensional analytical model; – with the help of BI tools, reporting services, and DW data mining, they enable the transformation of data into business information; – supporting decision-making based on OLAP analytical queries; – storage, aggregation, and analysis of data covering a wide time horizon;
		Limita-tions	– the need to plan analytical strategies before building a multidimensional model; – implementation of IT systems utilizing data warehouses is typically related to high costs and long implementation period;
Machine Learning	– input data, training, and test sets; – Machine learning algorithms	Opportunities	– improving decision-making by automating prioritization; – use of historical and incoming data; – data analysis using predictive algorithms; – task automation due to supporting business processes with IT systems using machine learning; – predicting the effects of a decision before making it.
		Limitations	– ineffective learning process on unstructured data; – possible interference in the operating of machine learning algorithms if the data contains incorrect values; – difficult cooperation between different machine learning algorithms;
Distributed Processing	– distributed file systems; – a platform for managing cluster resources; – framework enabling the implementation of the map and reducing operations.	Opportuni-ties	– possibility to store Big Data via a distributed file system; – possibility to process data from many sources with different structures; – reduction of Big Data processing time due to parallel calculations; – possibility to integrate with high-level programming languages (Python, Java);
		Limitations	– possible risks related to data security; – possible problems with working with large files due to restrictions of the distributed file system; – the need for manual implementation of algorithms by programmers; – possible delays when performing reduction and mapping operations; – required programming technical knowledge.

Source: Own work based on (*Apache MapReduce*, website state as of May 11, 2022; *What is machine learning*, website state as of May 11, 2022).

The indicated technologies constitute the basis of Big Data support in organizations.

4.4. Information management

Information management can be defined as “controlling the course of information processes in order to optimize them” (Bojar, 2002, p. 307). Specifically, information management concerns the creation, acquisition, organization, storage, distribution, and use of information (Detlor, 2010). The basic concepts related to information management are information systems, information processes, and information security. Information systems transform data into information using procedures involving an integrated environment of hardware, software, and people (Hasan, 2018). An information system integrates an organization and its processes, thereby increasing the efficiency of business management and its smooth operation. In the face of the challenges posed by Big Data, an efficient organizational information system should include processes, procedures, and technologies that will enable the transformation of data that meets the 5V attributes into valuable information necessary for the management and decision-making stage.

An organization’s information system should also implement information security requirements. Information processed in an organization constitutes its strategic resources. Security incidents targeting information assets are a company-wide threat. The effectiveness of an information system is determined, among other things, by information security. Information security in organizations should be based on the general theory of security and the theory of systems security (Szczepaniuk et al., 2020). Furthermore, the security of information resources should be considered in the context of possible and probable threats (Szczepaniuk, 2015). Guidelines regarding information security determine organizational and technical measures to provide it. For this reason, the authors postulate the positioning of the Information Security Management System (ISMS) in the proposed framework for the systemic processing of Big Data in an organization. ISMS aims to minimize the risk of information threats through planning, organizational, technical, and control activities (see e.g., Szczepaniuk et al., 2020). Information security management should consider legal regulations, human resources, technical solutions, and appropriate organizational and procedural solutions (see e.g., Szczepaniuk, 2016; Szczepaniuk, 2015).

Depending on the purpose, the cybersecurity management model can have many elements. Tvaronavičienė et al. (2021) identify seven elements of the cybersecurity management model for critical infrastructure, which relate to the following categories: organization, technology, cyberculture, law, security, and strategy.

5. Results of empirical research

The article presents the results of quantitative empirical research that allowed us to identify the determinants of the effectiveness of Big Data processing.

The empirical research was conducted using the diagnostic poll method, survey technique, and expert interview technique. A survey questionnaire and an interview questionnaire were used as research tools. In the statistical analysis of the empirical research results, the authors applied Pearson’s chi-square test of independence. A significance level of α = 0.05 was adopted. The following form of Pearson’s chi-square test was used (see e.g., Turhan, 2020):

$$\chi^2 = \sum_{j=1}^{k} \frac{\left(O_j - E_j\right)^2}{E_j} \tag{1}$$

key:
χ^2 – Pearson’s chi-square test,
O_j – observed number,
E_j – expected number.

In order to determine the strength of the identified relationships, the C-Pearson contingency coefficient was used (see e.g., Hartmann et al., 2018):

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}} \tag{2}$$

key:
C – the C-Pearson contingency coefficient,
χ^2 – the chi-square statistic,
N – the total sample size.

5.1. Structure of the processed data

The first stage of the research was to identify the structure of the data processed in the organizations under study. Recognition of the data structure included three key levels of analysis of the empirical research:
– identification of categories of Big Data sets processed in the surveyed organizations;
– identification of data models at the level of the organization's information systems;

The survey questionnaire formulated a series of questions about data processing structure in organizations. The results of the survey are presented in Table 4.

Table 4. Categories of Big Data processed in the surveyed organizations

Big Data category	always		often		sometimes		rarely		never		Σ
	N	%N	N	%N	N	%N	N	%N	N	%N	
unstructured data	29	16%	56	31%	60	34%	25	14%	8	5%	178
text data	21	12%	46	26%	56	31%	41	23%	14	8%	178
multimedia data	21	12%	46	26%	56	31%	41	23%	14	8%	178
flat files	43	24%	72	41%	33	19%	23	13%	6	3%	178
structured data	29	16%	61	34%	59	33%	22	12%	7	4%	178
geographic data	8	5%	12	7%	43	24%	68	38%	46	26%	178
real-time data	17	9%	31	17%	62	35%	50	28%	19	10%	178
natural language data	19	10%	29	16%	41	23%	60	34%	29	16%	178
linked data	12	7%	19	10%	31	17%	79	44%	37	21%	178

Source: Own work.

The table above shows that the surveyed organizations process various data sets. Respondents most often indicate the use of flat files and structured data. In turn, geographic data and data described in natural language are rarely processed. The processed data structure results directly from the specificity of the business processes performed in the surveyed organizations. However, the research indicates that organizations process both structured and unstructured data.

Effective processing of the categories of Big Data sets identified in Table 4 requires adequate data models at the level of information systems. Table 5 shows the primary data model implemented in the surveyed organizations' information systems for Big Data processing.

Table 5. Data models implemented in IT systems of the surveyed organizations

Implemented data model	Number of organizations (N)	%N
relational	61	34%
key-value	20	11%
document	30	17%
graph	21	12%
wide column	20	11%
object-oriented	26	15%
Σ	178	100%

Source: Own work.

Based on the results in Table 5, it can be concluded that the surveyed organizations mainly use the relational data model (34% of the surveyed companies) in processing Big Data. The article's theoretical part indicated possible limitations of the relational model in the studied dimension. NoSQL class solutions constitute the remaining part of the used solutions. The document data model, used by 17% of the surveyed companies, is the most popular NoSQL data model. The graph data model and the key-value data model are used the least frequently. The use of these data models as the main one in Big Data processing is indicated by 11% of the surveyed organizations.

5.2. Cybersecurity assessment of Big Data processing

In the literature on the subject, it is widely recognized that cybersecurity is related to three key attributes, i.e., confidentiality, availability, and integrity (see e.g., Aminzade, 2018; Samonas & Coss, 2014). The ISO/IEC 27001 standard also recommends the listed attributes at many stages of information security management (see e.g., ISO/IEC 27001). For this reason, our research is based on the information mentioned above security attributes to assess the cybersecurity of Big Data processing. For the purposes of the study, it was assumed that individual information security attributes mean:

– confidentiality – Big Data sets are readable only by authorized users and IT systems;

– availability – users and IT systems have access to Big Data resources at a defined time in order to implement specific processes;

– integrity – Big Data sets have not been accidentally or intentionally modified during processing, transmission, and storage, if not required.

The research assumed that the assessment would be based on the number of incidents related to the attributes mentioned above of information security. The results of the survey are presented in Table 6.

Table 6. Evaluation of cybersecurity of Big Data processing in organizations

Incidents related to information security attributes	very often		often		sometimes		rarely		very rarely or never		Σ
	N	%N	N	%N	N	%N	N	%N	N	%N	
Confidentiality	4	2%	7	4%	39	22%	89	50%	39	22%	178
Availability	12	7%	24	13%	94	53%	36	20%	12	7%	178
Integrity	18	10%	27	15%	88	49%	31	17%	14	8%	178

Source: Own work.

Research results show that most security incidents occur in the area of data integrity. 25% of the surveyed organizations indicate very frequent or frequent problems in ensuring the accuracy and completeness of Big Data sets. Similar results also apply to the availability attribute. On the other hand, security incidents in the field of confidentiality occur much less frequently. The research results show that integrity and availability are the key challenges of future Big Data cybersecurity strategies.

5.3. Big Data processing conditions

One of the empirical research's key objectives was to identify Big Data processing conditions in companies. This objective was achieved by:

– identification of technologies, techniques, and tools used in the surveyed organizations to process Big Data sets;

– identification of the most common problems related to Big Data processing.

The survey formulated a number of questions about the technologies and tools used in the surveyed organizations to process Big Data. The results of the research are presented in Table 7.

Table 7. Technologies and tools used to process Big Data in the surveyed organizations.

Technologies, tools	always		often		sometimes		rarely		never		Σ
	N	%N	N	%N	N	%N	N	%N	N	%N	
Inferential statistics	8	4%	25	14%	70	39%	58	33%	17	10%	178
Data warehouses	40	22%	68	38%	42	24%	22	12%	6	3%	178
Data mining	31	17%	64	36%	33	19%	35	20%	15	8%	178
Distributed processing	12	7%	28	16%	31	17%	33	19%	74	42%	178
Machine learning	8	4%	23	13%	28	16%	51	29%	68	38%	178
SQL queries	23	13%	41	23%	32	18%	38	21%	44	25%	178
Querying non-relational database systems	34	19%	43	24%	29	16%	26	15%	46	26%	178

Source: Own work.

Based on the obtained research results included in Table 7, it should be stated that the most popular technologies and tools used in Big Data processing are data warehouses and data mining. It means that the surveyed organizations rely on traditional IT solutions from the area of Business Intelligence. Some surveyed organizations point to the need to implement non-relational data models in Big Data processing based on implemented business processes. In terms of the least used technologies and tools, parallel and distributed processing and Machine learning should be pointed out. The research shows that the latest Big Data processing technologies are at the initial implementation stage in the surveyed organizations.

5.4. Big Data processing efficiency

As mentioned in the introduction, one of the article's aims was to identify the factors influencing the effectiveness of the organization's IT systems in Big Data processing. The evaluation of Big Data processing efficiency was carried out multidimensionally, considering the criteria resulting from information management. Based on the conducted theoretical analysis, the authors formulated a research hypothesis assuming that the effectiveness of Big Data processing in organizations depends on the data model implemented in information systems. In order to verify the hypothesis, we developed a research model in which a number of dependent and independent variables were distinguished. Dependent variables result from the distinguished criteria for assessing effectiveness based on the theoretical foundations of information management and are related to the guidelines for implementing the digital economy. In the developed research model, the following variables were distinguished:

– independent variable *A*: the data model implemented in the organization's information system;

– dependent variable *B*: efficiency of organizing information;

– dependent variable *C*: efficiency of information storage;

– dependent variable *D*: efficiency of producing new information.

The efficiency evaluation was initiated by identifying values for the assumed dependent variables. For this purpose, a number of questions were formulated in the survey questionnaire to evaluate the efficiency of Big Data processing in the surveyed organizations. Table 8 presents the obtained results of the research.

Table 8. The efficiency of Big Data processing in the surveyed companies

The efficiency evaluation criteria	high		average		low		Σ
	N	%N	N	%N	N	%N	
The efficiency of organizing information	52	29%	82	46%	44	25%	178
The efficiency of information storage	53	30%	89	50%	36	20%	178
The efficiency of producing new information	54	30%	79	44%	45	25%	178

Source: Own work.

To statistically verify the research hypothesis, the authors used Pearson's chi-square test of independence determined for all the distinguished criteria for assessing the efficiency of Big Data processing.

The study aimed to identify the relationship between independent variable *A* and dependent variables *B*, *C*, and *D*. The following set of statistical hypotheses was defined for the study:
- H_0 null hypothesis: there is no relationship between the independent and dependent variables;
- H_1 alternative hypothesis: independent and dependent variables are related to each other.

The value of the χ^2 independence test for variables *A* and *B* is 19.46, and the associated probability value p = 0.035 with 10 degrees of freedom. Since the probability value *p* is less than the assumed significance level $\alpha = 0.05$, the null hypothesis H_0 should be rejected in favor of the alternative hypothesis H_1. It can be concluded that the effectiveness of organizing information depends on the data model used.

The value of the χ2 independence test for variables *A* and *C* is 22.31, and the associated probability value $p = 0.01$ with 10 degrees of freedom. Since the probability value *p* is less than the assumed significance level $\alpha = 0.05$, the null hypothesis H_0 should be rejected in favor of the alternative hypothesis H_1. It can be concluded that the effectiveness of information storage depends on the data model used.

The value of Pearson's χ2 independence test for variables *A* and *D* is 20.53, and the associated probability value $p = 0.02$ with 10 degrees of freedom. Since the probability value *p* is less than the assumed significance level $\alpha = 0.05$, the null hypothesis H_0 should be rejected in favor of the alternative H_1. It can be concluded that the efficiency of creating new information depends on the data model used.

5.5. Verification of research hypotheses

The empirical research revealed statistically significant relationships between the dependent and independent variables. The C-Pearson contingency coefficient was used to determine the strength of the identified relationships. Table 9 presents the verification of the research hypotheses.

Table 9. Verification of the research hypotheses and strength of identified relationships

Dependent variables	Chi-2 test				C Pearson contingency coefficient	
	Empirical value	*p-value*	Significance level	Hypothesis verification	Empirical value	Relation strength
	Independent variable - data models (variable *A*)					
The efficiency of organizing information (variable *B*)	19,46	0,035	0,05	positive	0,82	strong
The efficiency of information storage (variable *C*)	22,31	0,01		positive	0,86	strong
The efficiency of producing new information (variable *D*)	20,53	0,02		positive	0,84	strong

Source: Own work.

The value of the C-Pearson contingency coefficient for all positively verified hypotheses is above 0.8, which indicates a strong relationship between the variables.

6. Model of effective Big Data processing

Empirical research has shown a statistically significant relationship between data models and the efficiency of Big Data processing in organizations. Based on the conducted literature review, theoretical analyses, and empirical findings, the authors propose extending the layered perspective proposed by Cox and Ellsworth (1997) with a synergistic approach that considers organizational and information systems requirements. To this end, a systemic, three-layer model for the systemic processing of Big Data in organizations has been developed (Figure 2).

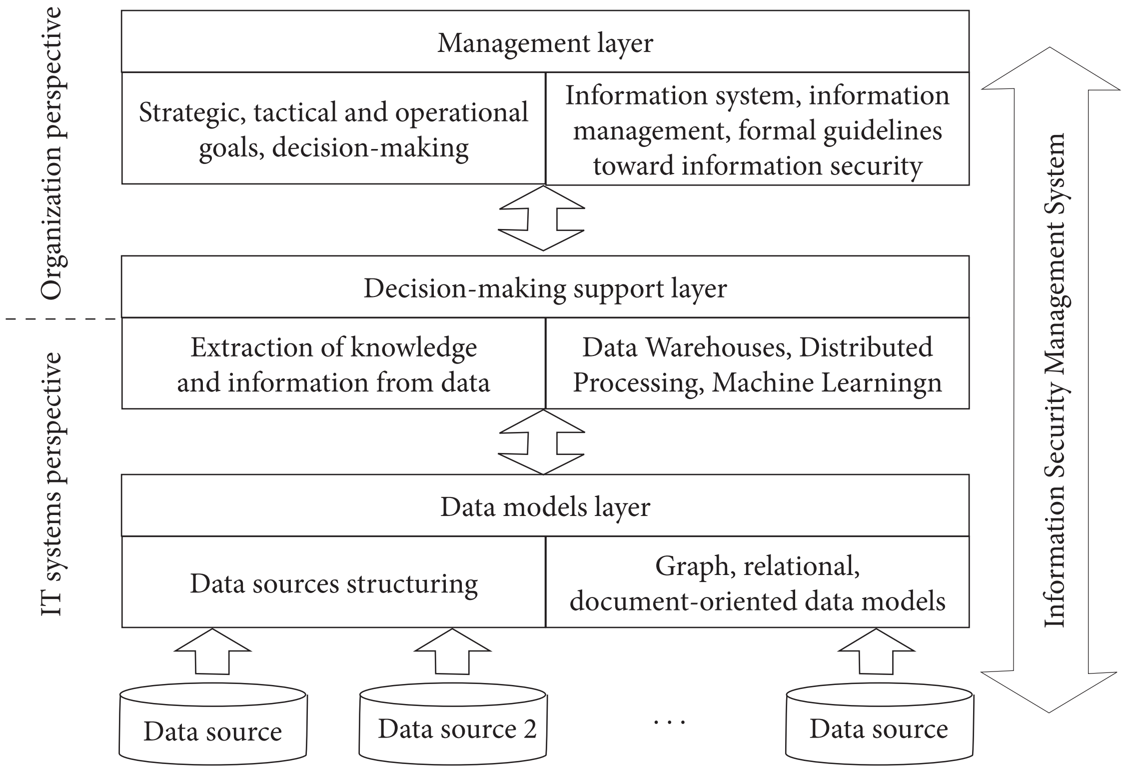

Figure 2. Model of systemic processing of Big Data in an organization

Source: Own work.

The proposed model includes:
– management layer,
– decision-making support layer,
– data models layer,
– cybersecurity of information resources.

The organization's operational and business activities generate data that meets the 5V attributes (volume, velocity, variety, veracity, and value). The indicated attributes constitute the entry to the proposed systemic Big Data processing model. The lowest layer of the proposed model is its foundation. In the layer of data models, data sources are structured for processing in higher model layers. The research shows that proper structuring of heterogeneous and inconsistent Big Data input is crucial for the effectiveness of decision support algorithms in higher layers of the model. Data that has been appropriately structured constitutes the input to the decision support layer. This layer contains algorithms and procedures for data analysis that enable information and knowledge extraction from data. The algorithms and procedures applicable in decision support are data warehousing, distributed processing, and Machine learning. Thanks to Big Data processing, the top layer is the beneficiary of the added value for the knowledge and information extracted in the lower layers. Simultaneously, the management layer is responsible for organizing the information system, managing information, defining performance criteria and guidelines for information security. Cybersecurity should be ensured at every stage of an organization.

The proposed systemic approach to Big Data in an organization fits into the primary areas of information management from the perspective of an organization and IT systems. Both perspectives are synergistically complementary, with the effectiveness of the proposed model being based on the resulting management decisions made on the basis of Big Data inputs. Based on the model shown in Figure 2, the output business knowledge obtained through Big Data processing by IT systems can be mathematically defined as follows:

$$\left.\begin{array}{lll} \bigwedge_{K\in\mathbb{K}} & \bigvee_{K_i} \rightarrow & K_i = f_i(A_i, I_i, S_i) \\ \bigvee_{I\in\mathbb{I}} & \bigvee_{I_i} \rightarrow & I_i = g_i(M_i, BD_i) \\ \bigvee_{S\in\mathbb{S}} & \bigvee_{S_i} \rightarrow & S_i = h_i(R_i, SM_i) \end{array}\right\} \quad K_i = f_i\{A_i, g_i(M_i, BD_i), h_i(R_i, SM_i)\} \qquad (3)$$

key:
K – business knowledge;
A – algorithms and procedures of the decision-making support layer;
I – business information;
S – cybersecurity of information resources;
M – models for structuring data sources from the data model layer;
BD – input Big Data resources;
R – cybersecurity requirements;
SM – security measures.

7. Conclusions

Research findings show that the identified factors of Big Data processing efficiency in organizations can support the implementation of international strategies in the digital economy and data-based business models. Based on the analysis of the obtained research results, a framework for effective Big Data processing in the organization was proposed, taking into account the identified efficiency factors.

Concerning the research goals, the following conclusions were formulated:

1) The technical aspects of designing IT systems, including software programming, support the implementation of international data strategies and the development of smart cities.

2) Barriers to the development of the digital economy and smart cities can be reduced by the Big Data processing efficiency factors identified in the study.

3) Data models are a critical intermediate layer between input data and data processing algorithms and determine the effectiveness of the entire Big Data analytics process.

4) Integrating distributed and heterogeneous Big Data resources can be reduced by models operating with a dynamic data schema (graph or document-oriented models).

5) The results of the cybersecurity evaluation indicate that the attribute of information integrity and availability will be the key challenges of future Big Data processing strategies.

6) Information security solutions should take into account potential threats and appropriate preventive measures. It is also crucial to implement the Information Security Management System, which will consider protecting strategic Big Data resources at every stage of the organization's functioning.

The issue of Big Data processing efficiency in organizations requires further research. The analysis of organizational processes suggests that the further development of Big Data technology will take place rapidly. Therefore, it is necessary to continue research on the effective processing of heterogeneous and inconsistent Big Data sets.

References

A European Strategy for data. An official website of the European Union. https://digital-strategy.ec.europa.eu/en/policies/strategy-data (accessed May 6, 2022).

Al-Jarrah, O.Y., Yoo, P.D., Muhaidat, S., Karagiannidis, G.K., & Taha, K. (2015). Efficient Machine Learning for Big Data: A Review. *Big Data Research*, 2(3), 87-93. https://doi.org/10.1016/j.bdr.2015.04.001

Aminzade, M. (2018). Confidentiality, integrity and availability – finding a balanced IT framework. *Network Security*. 2018(5), 9-11. https://doi.org/10.1016/S1353-4858(18)30043-6

Anthopoulos, L., & Kazantzi, V. (2022). Urban energy efficiency assessment models from an AI and big data perspective: Tools for policy makers. *Sustainable Cities and Society*, 76, Article 103492. https://doi.org/10.1016/j.scs.2021.103492

Apache MapReduce. An official website of the IBM. https://www.ibm.com/analytics/hadoop/mapreduce (accessed May 11, 2022).

Barham, H., & Daim, T. (2020). The use of readiness assessment for big data projects. *Sustainable Cities and Society*, 60, Article 102233. https://doi.org/10.1016/j.scs.2020.102233

Benzidia, S., Makaoui, N., & Bentahar, O. (2021). The impact of big data analytics and artificial intelligence on green supply chain process integration and hospital environmental performance. *Technological Forecasting and Social Change*, 165, Article 120557. https://doi.org/10.1016/j.techfore.2020.120557

Bhatia, S., Bagoria, R., & Kochar, B. (2019). A Literature Review on Quality Prediction in Data Warehouse. In *Proceedings of the International Conference on Advancements in Computing & Management,* Jaipur, India. http://dx.doi.org/10.2139/ssrn.3464938

Bi, H., & Ye, Z. (2021). Exploring ridesourcing trip patterns by fusing multi-source data: A big data approach. *Sustainable Cities and Society*, 64, Article 102499. https://doi.org/10.1016/j.scs.2020.102499

Big data. An official website of the European Union. https://digital-strategy.ec.europa.eu/en/policies/big-data (accessed May 6, 2022).

Boell, S.K., & Cecez-Kecmanovic, D. (2015). What is an Information System? *48th Hawaii International Conference on System Sciences*. IEEE. Kauai, HI, USA. https://doi.org/10.1109/HICSS.2015.587

Bojar, B. (2002). *Słownik encyklopedyczny informacji, języków i systemów informacyjno-wyszukiwawczych*. Warsaw: Stowarzyszenie Bibliotekarzy Polskich.

Chen, J., Ramanathan, L., & Alazab, M. (2020). Holistic Big Data Integrated Artificial Intelligent Modeling to Improve Privacy and Security in Data Management of Smart Cities. *Microprocessors and Microsystems,* 81, Article 103722. https://doi.org/10.1016/j.micpro.2020.103722

Codd, E.F. (1972). *Relational Completeness of Data Base Sublanguages.* San Jose: IBM Research Laboratory.

Cox, M., & Ellsworth, D. (1997). Managing Big Data for Scientific Visualization. Proc. ACM Siggraph. ACM, 97, 21-38.

Data Act: Businesses and citizens in favour of a fair data economy. An official website of the European Union. https://digital-strategy.ec.europa.eu/en/news/data-act-businesses-and-citizens-favour-fair-data-economy (accessed April 9, 2022).

DB-Engines Ranking. (2022). DB-Engines. https://db-engines.com/en/ranking (accessed June 5, 2022).

Detlor, B. (2010). Information management. *International Journal of Information Management,* 30(2), 103-108. https://doi.org/10.1016/j.ijinfomgt.2009.12.001

Dittrich, J., & Quiané-Ruiz, J.A. (2012). Efficient big data processing in Hadoop MapReduce. *Proceedings of the VLDB Endowment*, 5(12), 2014-2015. https://doi.org/10.14778/2367502.2367562

Esrafilian-Najafabadi, M., & Haghighat, F. (2021). Occupancy-based HVAC control using Deep Learning algorithms for estimating online preconditioning time in residential buildings. *Energy and Buildings*, 252, Article 111377. https://doi.org/10.1016/j.enbuild.2021.111377

European Commission. (2020). *Data governance and data policies at the European Commission.* Retrieved April 6, 2022 from https://ec.europa.eu/info/sites/default/files/summary-data-governance-data-policies_en.pdf

Fielt, E., Desouza, K. C., Gable, G., & Westerveld, P. (2019). Data-Driven Business Models and Professional Services Firms: A Strategic Framework and Transitionary Pathways. In: Xu, J., Zhu, B., Liu, X., Shaw, M., Zhang, H., Fan, M. (Eds). The Ecosystem of e-Business: Technologies, Stakeholders, and Connections. WEB 2018. Lecture Notes in Business Information Processing. Volume 357. Springer, Cham. https://doi.org/10.1007/978-3-030-22784-5_3

Goal 11: Make cities inclusive, safe, resilient and sustainable. United Nations Sustainable Development. https://www.un.org/sustainabledevelopment/cities/ (accessed May 19, 2022)

Günther, W.A., Rezazade Mehrizi, M.H., Huysman, M., & Feldberg, F. (2017). Debating big data: A literature review on realizing value from big data. *The Journal of Strategic Information Systems,* 26(3), 191-209. https://doi.org/10.1016/j.jsis.2017.07.003

Hartmann, K., Krois, J., Waske, B. (2018). E-Learning Project SOGA: Statistics and Geospatial Data Analysis. Department of Earth Sciences, Freie Universitaet Berlin. https://www.geo.fu-berlin.de/en/v/soga/Basics-of-statistics/Descriptive-Statistics/Measures-of-Relation-Between-Variables/Contingency-Coeficient/index.html

Hasan, F.F. (2018). A Review Study of Information Systems. *International Journal of Computer Applications,* 179(18), 15-19. https://doi.org/10.5120/ijca2018916307

ISO/IEC 27001:2017, Information technology – Security techniques – Information security management systems – Requirements.

Jenkin, T.A., Chan, Y.E., & Sabherwal, R. (2019). Mutual Understanding in Information Systems Development: Changes Within and Across Projects. *MIS Quarterly,* 43(2), 649–671. https://doi.org/10.25300/MISQ/2019/13980

Karatas, M., Eriskin, L., Deveci, M., Pamucar, D., & Garg, H. (2022). Big Data for Healthcare Industry 4.0: Applications, challenges and future perspectives. *Expert Systems with Applications*, 200, Article 116912. https://doi.org/10.1016/j.eswa.2022.116912

Khan, M., Babar, M., Ahmed, S.H., Shah, S.C., & Han, K. (2017). Smart city designing and planning based on Big Data Analytics. *Sustainable Cities and Society*, 35, 271–279. https://doi.org/10.1016/j.scs.2017.07.012

Kisielicki, J. (2009). *Systemy informatyczne zarządzania*. Warsaw: PWN.

Koman, G., Bubelíny O., Tumová, D., Jankal, R. (2022). Sustainable transport within the context of smart cities in the Slovak republic. *Entrepreneurship and Sustainability Issues*, 10(1), 175-199. http://doi.org/10.9770/jesi.2022.10.1(9)

Korneć, R. (2020). The role of stakeholders in shaping smart solutions in Polish cities. *Entrepreneurship and Sustainability Issues,* 7(3), 1981-1995. https://doi.org/10.9770/jesi.2020.7.3(36)

Krishnamoorthy, M., Ahamed B., B., Suresh, S., & Alagappan, S. (2020). Deep Learning Techniques and Optimization Strategies in Big Data Analytics: Automated Transfer Learning of Convolutional Neural Networks Using Enas Algorithm. In: Thomas, J.J., Karagoz, P., Ahamed, B.B., and Vasant, P. (Eds). Deep Learning Techniques and Optimization Strategies in Big Data Analytics. Hershey, PA: IGI Global, pp. 142-153. https://doi.org/10.4018/978-1-7998-1192-3.ch009

Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity, and Variety. Application Delivery Strategies. Stamford: META Group Inc.

Laudon, K.C., & Laudon, J.P. (2012). *Management Information Systems: Managing the Digital Firm*. 12th Edition. Upper Saddle River, NJ: Prentice Hall.

Limba, T., Novikovas, A., Stankevičius, A., Andrulevičius, A., & Tvaronavičienė, M. (2020). Big Data Manifestation in Municipal Waste Management and Cryptocurrency Sectors: Positive and Negative Implementation Factors. *Sustainability*. 12(7), Article 2862. https://doi.org/10.3390/su12072862

Lin D., Cai L., Zhang X., Zhang X., & Huo J. (2020). Latency Estimation of Big Data Processing Under the MapReduce Framework with Coupling Effects. In: Liang, Q., Wang W., Mu, J., Liu, X., Na, Z., and Chen, B. (Eds). *Artificial Intelligence in China. Lecture Notes in Electrical Engineering*. Singapore: Springer, pp. 105-112. https://doi.org/10.1007/978-981-15-0187-6_12

Lomotey, R.K., & Deters, R. (2014). Towards Knowledge Discovery in Big Data. *2014 IEEE 8th International Symposium on Service Oriented System Engineering*. Oxford, UK, pp. 181-191. https://doi.org/10.1109/SOSE.2014.25

Loebbecke, C., & Picot, A. (2015). Reflections on societal and business model transformation arising from digitization and big data analytics: A research agenda. *The Journal of Strategic Information Systems*, 24(3), 149-157. https://doi.org/10.1016/j.jsis.2015.08.002

Ma, D., Hu, J., & Yao, F. (2021). Big data empowering low-carbon smart tourism study on low-carbon tourism O2O supply chain considering consumer behaviors and corporate altruistic preferences. *Computers & Industrial Engineering*, 153, Article 107061. https://doi.org/10.1016/j.cie.2020.107061

Maitrey, S., & Jha, C.K. (2015). MapReduce: Simplified Data Analysis of Big Data. *Procedia Computer Science*, 57, 563-571. https://doi.org/10.1016/j.procs.2015.07.392

Obermeyer, Z., & Emanuel, E.J. (2016). Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. *The New England Journal of Medicine*, 375, 1216-1219. https://doi.org/10.1056/NEJMp1606181

Qiu, J., Wu, Q., Ding, G., Xu, Y., & Feng, S. (2016). A survey of machine learning for big data processing. *EURASIP Journal on Advances in Signal Processing*, 2016, Article 67. https://doi.org/10.1186/s13634-016-0355-x

Panwar, A., & Bhatnagar, V. (2020). Architektura Data Lake: A New Repository for Data Engineer. *International Journal of Organizational and Collective Intelligence,* 10(1), 63-75. https://doi.org/10.4018/IJOCI.2020010104

Pustokhin, D.A., Pustokhina, I.V., Rani, P., Kansal, V., Elhoseny, M., Joshi, G.P., & Shankar, K. (2021). Optimal deep learning approaches and healthcare big data analytics for mobile networks toward 5G. *Computers & Electrical Engineering*, 95, Article 107376. https://doi.org/10.1016/j.compeleceng.2021.107376

Samonas, S., & Coss, D. (2014). The CIA Strikes Back: Redefining Confidentiality, Integrity and Availability in Security. *Journal of Information System Security*, 10(3), 21-45.

Scabora, L., Brito, J., Ciferri, R., & Ciferri, C. (2016). Physical Data Warehouse Design on NoSQL Databases - OLAP Query Processing over HBase. In: *Proceedings of the 18th International Conference on Enterprise Information Systems (ICEIS 2016).* Rome, Italy.

Stefanowski, J. (2009). Wprowadzenie do maszynowego uczenia się. Instytut Informatyki. Politechnika Poznańska. Retrieved May 20, 2022 from http://www.cs.put.poznan.pl/jstefanowski/ml/NEWML_W1lastr.pdf

Stergiou, C.L., Plageras, A.P., Psannis, K.E., & Gupta, B.B. (2020). Secure Machine Learning Scenario from Big Data in Cloud Computing via Internet of Things Network. In: Gupta, B., Perez, G., Agrawal, D., & Gupta, D. (Eds) *Handbook of Computer Networks and Cyber Security*. Cham: Springer, pp. 525-554.

Szczepaniuk, E., & Szczepaniuk, H. (2019). Effectiveness of Selected Database Models in Management Information Systems. *Organization and Management*, 1(45), 77-85. https://doi.org/10.29119/1899-6116.2019.45.8

Szczepaniuk, H. (2017). Zastosowania grafowych baz danych w systemach informatycznych zarządzania. *Ekonomika i organizacja logistyki*, 2(3), 85-94.

Szczepaniuk, E. (2015). Zarządzanie bezpieczeństwem informacji w administracji publicznej. In: Sienkiewicz, P. (Ed.) *Inżynieria systemów bezpieczeństwa*. Warsaw: Polskie Wydawnictwo Ekonomiczne, pp. 80-94.

Szczepaniuk, E. (2016). *Bezpieczeństwo struktur administracyjnych w warunkach zagrożeń cyberprzestrzeni Państwa*. Warszawa: Wydawnictwo Akademii Obrony Narodowej.

Szczepaniuk, E., Szczepaniuk, H., Rokicki, T., & Klepacki B. (2020). Information security assessment in public administration. *Computers & Security*. Volume 90. Article number 101709. https://doi.org/10.1016/j.cose.2019.101709

Szczepaniuk, H. (2017). Zastosowania grafowych baz danych w systemach informatycznych zarządzania. *Ekonomika i organizacja logistyki*, 2(3), 85-94.

Turhan, N.S. (2020). Karl Pearson's chi-square tests. *Educational Research and Reviews*, 15(9), 575-580. https://doi.org/10.5897/ERR2019.3817

Tvaronavičienė, M., Plėta, T., & Della Casa, S. (2021). Cyber Security Management Model for Critical Infrastructure Protection. In Proceedings of the Selected papers of the International Scientific Conference *Contemporary Issues in Business, Management and Economics Engineering,* May 13-14, 2021, Vilnius, Lithuania.

Ullah, F., Habib, M.A., Farhan, M., Khalid, S., Durrani, M.Y., & Jabbar, S. (2017). Semantic interoperability for big-data in heterogeneous IoT infrastructure for healthcare. *Sustainable Cities and Society*, 34, 90-96. https://doi.org/10.1016/j.scs.2017.06.010

Ullah, F., Qayyum, S., Thaheem, M. J., Al-Turjman, F., & Sepasgozar, S. M. E. (2021). Risk management in sustainable smart cities governance: A TOE framework. *Technological Forecasting and Social Change*, 167, Article 120743. https://doi.org/10.1016/j.techfore.2021.120743

Xing, E.P., Ho, Q., Dai, W., Kim, J.K., Wei, J., Lee, S., Zheng, X., Xie, P., Kumar, A., & Yu, Y. (2015). Petuum: A New Platform for Distributed Machine Learning on Big Data. *IEEE Transactions on Big Data*, 1(2), 49-67. https://doi.org/10.1109/TBDATA.2015.2472014

Vu, K., Hanafizadeh, P., & Bohlin, E. (2020). ICT as a driver of economic growth: A survey of the literature and directions for future research. *Telecommunications Policy*, 44 (2), Article 101922. https://doi.org/10.1016/j.telpol.2020.101922

What is machine learning? An official website of the SAP. https://www.sap.com/insights/what-is-machine-learning.html (accessed May 11, 2022).

Zhou, L., Pan, S., Wang, J., & Vasilakos, A.V. (2017). Machine learning on big data: Opportunities and challenges. *Neurocomputing*, 237, 350-361. https://doi.org/10.1016/j.neucom.2017.01.026

Zhao, F., Prentice, C., Wallis, J., Patel, A., Waxin, M.-F. (2020). An integrated study of the implications of the rise of coworking spaces in smart cities. *Entrepreneurship and Sustainability Issues*, 8(2), 467-486. http://doi.org/10.9770/jesi.2020.8.2(28)

ISSN 2029-7017 print/ISSN 2029-7025 online

Hubert SZCZEPANIUK is a researcher at the Warsaw University of Life Sciences. He is a graduate of bachelor's, master's and doctoral studies in the field of computer science. He also completed doctoral studies in social sciences and postgraduate studies in the area of pedagogy and mathematics. Research interests apply to computer science, cybersecurity, and management. His research activities also include the development of software in various technology stacks and the implementation of cybersecurity architectures and machine learning systems.
ORCID ID: https://orcid.org/0000-0001-7613-4281

Edyta Karolina SZCZEPANIUK is a researcher at the Polish Air Force University. Doctor of social sciences in the field of security sciences. She is a graduate of computer science, national security, and administration. Her scientific interests include information security, cybersecurity, e-administration, management information systems, and programming of systems and databases. Author of numerous publications in this field.
ORCID ID: https://orcid.org/0000-0002-6707-2987