Научная статья / Research article

Digital Transformation and Big Data

Stanislav S. Kuzora[1] ✉, **Ivan P. Natarov[2]**

[1]Far Eastern Federal University,
8 Sukhanova str., Vladivostok, Russian Federation, 690091
[2]Ministry of Economic Development of the Primorsky Region,
22 Svetlanskaya str., Vladivostok, Russian Federation, 690110
✉ webkuzora@mail.ru

Abstract. Today, the Russian Federation pays special attention to the digital transformation of the country as a whole and shows a separate interest in big data in particular, which is confirmed by a number of analyzed documents and the current situation in the subject area. The research work proposed for reading is devoted to the analysis of big data, one of the areas of use of which is the state, or rather the work of federal and regional executive authorities. The main value of big data from the position of the state is seen in the possibility of working on volumes of heterogeneous information in order to increase efficiency in making managerial decisions on a wide range of issues. The main purpose of the study is to disseminate the experience of big data analysis, which is used in the work of the Ministry of Economic Development of Primorsky Krai. To do this, it is proposed to turn to the theoretical aspects of big data, find out the main chain of work on them, and also pay attention to domestic and foreign experience in using data in some areas of knowledge. The practical part of the study is a description of the experience of the Ministry of Economic Development of the PC, which is directly related to the digital development of the region and carries out a certain list of works with heterogeneous data. As an example, attention is drawn to a proven approach to the analysis of a large set of open data characterizing the work and development of small and medium-sized businesses in the Primorsky Territory, posted on the website of the Federal Tax Service of the Russian Federation. In conclusion, it is emphasized that the approach proposed in the study for working with data can be adapted to similar needs in other federal and regional authorities

Keywords: Big Data, data analysis, digital transformation, government and Big Data, Big Data practice

Conflicts of interest: The authors declared no conflicts of interest.

Introduction

Before we talk about such a phenomenon as big data, let's pay attention to the enlarged field of knowledge "Data Science", which is an interdisciplinary field based on computer science and mathematics, applicable for data mining, machine learning and working with big data. Data science is both an academic and a practical discipline of the intersectoral sphere of human activity, aimed at studying digital data in order to extract useful information.

In the theoretical part of this paper, it is proposed to focus on big data, namely on the main stages of work on them: generation, collection, storage and analysis. At the same time, the general situation regarding the subject area will be considered on the basis of the analysis carried out. At the end of the section, a transition is made from the use of big data in many fields of activity to the application of a specific data set in a specific area. We hope that this way of presenting information will allow the reader to get a general idea of big data and pay attention to a case of their use.

Theoretical aspects of Big Data

The term "Big Data" is used to refer to structured, semi-structured and unstructured data volumes of a wide variety [1; 2]. The existing diversity depends on the field of knowledge. According to the filters of the Scopus bibliographic database, big data covers 27 branches of knowledge. The first five areas, which include computer technology, engineering, mathematics, social sciences and the science of decision-making, are most associated with big data. It may follow from this that several certain methods and tools can be used to work on the data of the designated areas [3], while other areas, less related to a large amount of data, provide for the use of other means of information processing.

Another feature of big data is a set of "three V" features (volume, speed, diversity), which has been developed to five, in some cases up to seven features [4]. These features emphasize the importance of not only the volume of big data, but also other essential characteristics when working with certain

data sets. In continuation, it is important to note that data volumes change rapidly, grow over time, and work on them changes with the development and emergence of new technologies.

Today, as a rule, big data is measured in terabytes, petabytes, exabytes and zettabytes. According to McKinsey & Company estimates, by 2009, many sectors of the US economy had on average at least 200 terabytes of stored data (twice as much as the data warehouse of the American retail chain Walmart in 1999) per company with more than 1,000 employees [5].

Today, society's attitude to data has changed with the development of multimedia, social networks, and the Internet of Things [6]. Now data has become one of the important production factors for enterprises and businesses [7; 8]. Recently, many government agencies around the world have joined the research and practical application of big data [9]. They can also be used by one person in the process of intellectual analysis and analytical processing for decision-making.

In addition to the traceable advantages of big data, working on them is accompanied by a number of problems. For example, one of the main ones is the difficulty in detecting data sets with a large scale, different types and rapid generation with subsequent extraction of valuable information [10]. Another example of complexity is the continuous development of data processing tools and the emergence of new technologies designed for high-speed collection, storage and analysis, which leads to additional financial and other costs for the end user [11].

After a preliminary acquaintance with big data, it is proposed to pay attention to the main chain of work on them, consisting of four stages.

1) *Generation* is the first step in working with data, aimed at finding information in the form of records, files, messages in the area of interest. Such data has different sources, format, volume and other distinctive characteristics. Individually, they may not be of any value, but their totality allows you to identify useful information on the basis of which you can get answers to questions of interest, track various kinds of patterns, find out people's habits, get other kinds of information.

2) *Collection* is the next step in working with data, including the process of transmission and preprocessing. During data collection with the help of certain means/tools, a huge amount of raw information accumulates, which may be redundant or useless. After collecting the raw data, the information is transferred through special applications to the appropriate storage. To reduce information redundancy, preprocessing is carried out by means of data compression technology and the participation of a specialist, which makes it possible to efficiently store and use data.

3) *Storage* is the process of managing preprocessed data, ensuring the reliability and availability of information. Data storage devices are relational database management systems (DBMS) (for example, PostgreSQL), NoSQL databases, etc. Such software systems are often used for a wide class of data storage tasks, on the basis of which it is possible to develop unique applications for the purpose of constant interaction with the data sets of interest. With the increased growth of global data volumes, the suppliers of the above-mentioned DBMS are constantly increasing the maximum allowable storage capacity, which indicates a high need for additional data storage space on the part of users.

4) *Analysis* is the most important link when working with big data, since its result is intended to give some value to the end user. Before the advent of modern methods of data analysis, statistical methods of analyzing heterogeneous data arrays were used. The most common are regression, correlation, and cluster data analysis, which through digitalization have become applicable through programming. This has led to the emergence of separate libraries, frameworks in certain programming languages, such as Payton, which combine knowledge of statistics and computer science. Based on the above, such capabilities can be considered a modern way of analyzing big data.

As a summary of this section, it should be noted that big data analysis plays a huge role in understanding the needs of business customers, forecasting market trends of large organizations, and developing development strategies in many countries of the world. As can be seen, big data analysis is a rather laborious process that requires special knowledge and effort for practical application. In the next section, it is proposed to pay attention to the role of big data in the context of the digital transformation of the Russian Federation, and then consider one of the possible ways to analyze data in the process of the work of a regional executive authority.

Digital
public administration

Particular attention has been paid to the development of the digital economy of the Russian Federation at the legislative level since the formation of the national program "Digital Economy of the Russian Federation", approved by the minutes of the meeting of the Presidium of the Council under the President of the Russian Federation on Strategic Development and National Projects dated June 4, 2019, No. 7.

Within the framework of the national projects of Russia formed by the Government of the Russian Federation, federal and regional executive

authorities have initiated such projects as: digital public administration, personnel for the digital economy, digital technologies, etc. [12-14]. These projects are aimed at accelerating the introduction of digital technologies in order to create conditions for high-tech business, increase the competitiveness of the country, strengthen national security and improve the quality of life of citizens.

Based on the analyzed documents, approved strategies, planned projects, it is noticeable that much attention is paid to digital data, which cover many aspects of socio-economic activity in the field of public administration, education, industry and business [15]. Despite the rather wide list of the above-mentioned areas that are supposed to be the basis for the digital transformation of Russia, within the framework of this work it is proposed to focus on such a direction as "Big Data", namely, the use of big data in the activities of the Ministry of Economic Development of Primorsky Krai.

According to the regulatory framework, the use of big data, namely the extraction of the necessary information from them, is mainly aimed at the implementation of regulatory documentation in the public sector, the development of new educational programs, the formation of appropriate competencies when working with data, the creation of technological foundations for the implementation of new projects. It is also noted that for high-quality management of digital development, big data can be used as a tool to increase transparency of the activities of all stakeholders.

Let's pay attention to a special case of the need to use big data as part of digital transformation of certain sectors of the Primorsky Krai economy. The main document of the transformation is the strategy, the purpose of which is to increase the digital maturity of the socio-economic sphere in certain areas, including public administration. On the part of public administration, the strategy identifies a number of specific tasks, the implementation of which is hardly possible without the use of new technologies and work with big data.

Moreover, the document assigns a special role to the problems and challenges that will have to be faced in the process of digital transformation, where the use of big data is associated, for example, with paragraph 4.5.8 "Lack of a single information space for storing, consolidating and processing electronic data to ensure the timeliness and reliability of providing information to executive authorities".

The strategy also lists a list of planned projects, required resources and established indicators to ensure the implementation of the digital transformation of the region. The responsible executive authority for digital public administration is the Ministry of Digital Development and

Communications of Primorsky Krai. However, other public administration entities, for example, the Ministry of Economic Development of Primorsky Krai, also set strategic guidelines for the transformation of the digital environment.

To date, the mentioned bodies have formed work plans and activities that include the use of information technologies, including new methods and tools when working with different data sets. In view of the rather early period of digital transformation, especially in the field of public administration, which is confirmed by regulatory and legal documentation and the lack of implementation of completed projects in the public information space, work with big data is at the initial stage of formation. Despite the short period of development of this direction, an example of working with certain data sets already exists.

Big Data experience

As mentioned earlier, the Ministry of Economic Development of the PC is directly related to the digital development of the region and carries out a certain list of works with data. In this regard, it is proposed to pay attention to the proven approach to the analysis of a large set of open data characterizing the work and development of small and medium-sized businesses in Primorsky Krai, posted on the website of the Federal Tax Service of the Russian Federation.

For the convenience of familiarization with the material, the structure of the section has the following form: designation of relevance, setting goals and objectives of the analysis, description of the chain of work on the data, summing up and recommendations.

Relevance. One of the tasks of the Ministry of Economic Development of Primorsky Krai is the analysis of the socio-economic situation in the region, the formation of regional development mechanisms, the adoption of balanced and informed decisions aimed at the development of the region. Today, in the public sector, various kinds of information systems generate petabytes of data annually, which are often not used by government officials, but only placed in open sources for external users.

As an example, we can single out the sphere of small and medium-sized businesses. On a monthly basis, the Federal Tax Service (hereinafter referred to as the FTS) generates gigabytes of information on the development of entrepreneurship both in the regions and in the Russian Federation on the basis of the Unified Register of Small and Medium–Sized Businesses and

other data sources. Due to the specific structure of the data, their analysis is complicated and requires technical skills to process and bring it into an understandable form.

Purpose and objectives. The purpose of the work from the point of view of the formulated relevance is to obtain generalizing indicators and identify the level of tax solvency of small and medium-sized businesses in the territories of municipalities Primorsky Krai. To achieve this goal, it is necessary to solve the following tasks: 1) data collection and preprocessing, 2) data set formation, 3) statistical data analysis.

Working on data.

The first stage (data collection and preprocessing). The following data describing the development of small and medium-sized businesses in Primorsky Krai is being uploaded from the official website of the Federal Tax Service:
1. Data from the Unified Register of Small and Medium-sized Businesses.
2. Information on the amounts of income and expenses according to the accounting (financial) statements of the organization for the year.
3. Information about special tax regimes applied by taxpayers.
4. Information on the amounts of arrears and arrears on penalties and fines.
5. Information about the amounts of taxes and fees paid by the organization in the calendar year.
   As a result, 54,481 files were uploaded, of which:
   - 6,211 files fall on the melons of the Unified Register of Small and Medium-sized Businesses
   - 12,063 files contain information about income and expenses;
   - 12,067 files contain information about special tax regimes;
   - 12,056 files contain information about the amounts of arrears;
   - 12 084 files contain information about taxes paid.

All data is in XML format, which leads to the impossibility of subsequent analysis, and as a result, data is preprocessed by programming languages.

The second stage (formation of a data set). With the help of the Python programming language, scripts were written for processing the received data, followed by their conversion to CSV (Comma-Separated Values) format, convenient for use by Python libraries (Pandas) in a tabular representation.
   As a result, the following data sets are obtained:
   - reestr_msp_2019.csv;
   - company_revenue_2019.csv;
   - company_tax_regime_2019.csv;

- msp_arrears_2019.csv;
- msp_tax_2019.csv.

Further, the data obtained were combined into a final data set (data_msp.csv) for subsequent statistical analysis. After combining, the final set consists of 38,366 lines and 53 attributes, where each line characterizes one legal entity that has the attribute of a small and medium–sized business entity, and the attributes (columns) are the characteristics of this entity, for example, the type of activity carried out, the business category (micro, small or medium), the applicable taxation system, revenue, taxes paid, tax arrears, etc.

The third stage (statistical data analysis). The final stage of work on the data is statistical analysis, which includes research, correlation analysis and hypothesis testing. In order to avoid excessive description of this stage, it is proposed to briefly explain its constituent elements.

Research data analysis (EDA — exploratory data analysis) provides:

- data description (study of data dimension, how many rows and columns are in the data, description of data types, study of measures of central trend and data distribution);
- changing data types (converting non-numeric values into a categorical form, converting a date from a string value to a date format, etc.);
- description of omissions (identification of the percentage of missing values in the data and their subsequent deletion or restoration);
- description of duplicates (search for duplicate information in the data);
- work with categorical features (spell checking, standardization of names, if necessary, etc.);
- conclusion (the result of calculations, display of graphs, construction of a correlation matrix, etc.).

For subsequent analysis, two hypotheses were formulated.

Hypothesis No. 1. Municipalities differ in the level of tax solvency of small and medium-sized businesses in Primorsky Krai.

Hypothesis No. 2. There is a group of municipalities Primorsky Krai, which has the lowest level of tax solvency among small and medium-sized businesses of Primorsky Krai.

Hypothesis testing was carried out using the Kraskel-Wallis criterion and as a result, the following conclusions were obtained:

- municipalities of Primorsky Krai differ in the level of tax solvency of small and medium-sized businesses;
- there are groups of municipalities Primorsky Krai, having the lowest level of tax solvency among small and medium-sized businesses of Primorsky Krai.

A link to the full analytical report on the work done is provided in the list of sources used [16].

Results and recommendations. Based on the results of the work done, the goal of obtaining generalizing indicators and identifying the level of tax solvency among small and medium-sized businesses in Primorsky Krai has been achieved. It is established that municipalities differ in the level of tax solvency, and there is a group of municipalities with the lowest level of tax solvency.

As recommendations, it can be noted that infrastructure organizations supporting small and medium-sized businesses in Primorsky Krai, when forming an annual plan of business training activities on tax literacy, need to pay attention to municipalities with a low level of tax solvency.

Another recommendation may be addressed to the control and supervisory authorities with powers in the field of tax control, in order to increase the level of tax receipts to the consolidated budget of Primorsky Krai, it is necessary to strengthen tax control in the territories of municipalities.

Conclusion

Summing up the results of the study, it is proposed to focus on the key conclusions and features of the work done:

- big data is an integral part of the digital economy of the Russian Federation, the competent use of which can lead to the achievement of national goals and the solution of strategic tasks;
- implementation of such federal and regional projects as "Digital Public Administration", as well as strategies in the field of digital transformation is inextricably linked with work on heterogeneous data;
- familiarity with the approach of working with certain data sets with the participation of the Ministry of Economic Development of Primorsky Krai showed one of the possible options for analyzing unstructured data.
- the proposed approach has the potential to adapt to work with different data sets in the fields of activity of both the Ministry of Economic Development of the Russian Federation and other federal and regional executive authorities of the Russian Federation.

REFERENCES

1. Chen M., Mao S., Liu Y. Big Data: A Survey. *Mobile Netw Appl*, 2014: 171–209. DOI: 10.1007/s11036-013-0489-0

2. Longbing C. Data Science: A Comprehensive Overview. *ACM Comput. Surv*. 2017;50(3):42. DOI: https://doi.org/10.1145/3076253

3. Zvyagin L.S. Cifrovye trendy v analize dannyh i mjagkih izmerenijah kak konceptual'naja osnova razvitija prikladnyh nauk [Mathematical Algorithms of Game Theory as an Applied Tool for Making Effective Financial and Economic Decisions]. *Myagkiye izmereniya i vychisleniya*. 2020;37(12):45–62 (In Russ.).

4. Chen J., Jiang Q., Wang Y., Tang J. Study of Data Analysis Model Based on Big Data Technology. *2016 IEEE International Conference on Big Data Analysis (ICBDA)*; 2016: 1–6. DOI: https://doi.org/10.1109/ICBDA.2016.7509810

5. Big Data: The Next Frontier for Innovation, Competition, and Productivity. *McKinsey Global Institute.* URL: https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation#. Accessed: 23.01.2022

6. Platonova S.I. Big Data: Challenges and Opportunities in Social Sciences. *Manuskript*. 2020; 13(4). DOI: https://doi.org/10.30853/manuscript.2020.4.24 (In Russ.).

7. Chernyak L. Bol'shiye Dannyye — novaya teoriya i praktika [Big Data — New Theory and Practice]. *Otkrytyye sistemy. SUBD*. 2011;10:18–25 (In Russ.).

8. Chernyak L. Svezhiy vzglyad na Bol'shiye Dannyye [A New Look at Big Data]. *Otkrytyye sistemy. SUBD*. 2013;7:48–51 (In Russ.).

9. Volkov D.V. V poiskakh sokrovishch [Looking for Treasures]. *Otkrytyye sistemy. SUBD*. 2014;1:1 (In Russ.).

10. Sagiroglu S., Sinanc D. Big Data: A Review. *2013 International Conference on Collaboration Technologies and Systems (CTS)*; 2013:42–47. DOI: https://doi.org/10.1109/CTS.2013.6567202

11. Philip Chen C.L., Chun-Yang Z. Data-Intensive Applications, Challenges, Techniques and Technologies: A Survey on Big Data. *Information Sciences*. 2014;275:314–347. DOI: https://doi.org/10.1016/j.ins.2014.01.015

12. *Regional'nye proekty cifrovoj jekonomiki Primorskogo* kraja [Regional Projects of the Digital Economy of Primorsky Region]. URL: https://digital.primorsky.ru/. Accessed: 23.01.2022 (In Russ.).

13. *Ukaz Prezidenta RF ot 09.05.2017 N 203 «O Strategii razvitija informacionnogo obshhestva v Rossijskoj Federacii na 2017 — 2030 gody»* [Decree of the President of the Russian Federation of May 9, 2017 # 203 "On the Strategy for the Development of the Information Society in the Russian Federation for 2017 — 2030"]. URL: https://base.garant.ru/71670570/. Accessed: 23.01.2022 (In Russ.).

14. *Ukaz Prezidenta RF ot 21.07.2020 N 474 "O nacional'nyh celjah razvitija Rossijskoj Federacii na period do 2030 goda"* [Decree of the President of the Russian Federation of July 21, 2020 # 474 "On the National Development Goals of the Russian Federation for the Period up to 2030"]. URL: http://publication.pravo.gov.ru/Document/View/0001202007210012. Accessed: 23.01.2022 (In Russ.).

15. *Strategija v oblasti cifrovoj transformacii otraslej jekonomiki, social'noj sfery i gosudarstvennogo upravlenija Primorskogo kraja* [Strategy in the Field of Digital Transformation of Sectors of the Economy, Social Sphere and Public Administration of Primorsky Region]. URL: https://www.tadviser.ru/images/7/7c/Стратегия_Приморский_край.pdf. Accessed: 23.01.2022 (In Russ.).

16. *Proekt po analizu nalogovoj platezhesposobnosti malogo i srednego biznesa na territorijah municipal'nyh obrazovanij Primorskogo kraja* [Project to Analyze the Tax Solvency of Small and Medium-sized Businesses in the Territories of Municipalities of Primorsky Region]. URL: https://github.com/ivannatarov/Data_analysis_msp. Accessed: 23.01.2022 (In Russ.).

Information about the authors:

Stanislav S. Kuzora — Assistant of the Department of Innovations, Polytechnic Institute, Far Eastern Federal University, Russian Federation, ORCID: 0000-0003-2309-2035, e-mail: kuzora_ss@dvfu .ru

Ivan P. Natarov — Data Analyst, Ministry of Economic Development of the Primorsky Region, Russian Federation, ORCID: 0000-0002-6550-6833, e-mail: natarov_ivan@bk.ru

Цифровая трансформация и большие данные

С.С. Кузора[1] ⬤ ✉, И.П. Натаров[2] ⬤

[1]Дальневосточный федеральный университет,
690091, Россия, Владивосток, ул. Суханова, 8
[2]Министерство экономического развития Приморского края,
690110, Россия, Владивосток, ул. Светланская, 22
✉webkuzora@mail.ru

Аннотация. На сегодняшний день Российская Федерация уделяет особое внимание цифровой трансформации страны в целом и проявляет отдельный интерес к большим данным в частности, что подтверждается рядом проанализированных документов и сложившейся конъюнктурой в предметной области. Исследование посвящено анализу больших данных, одной из областей использования которых является государство, а точнее работа федеральных и региональных органов исполнительной власти. Основная ценность больших данных с позиции государства видится в возможности работы над объемами разнородной информации с целью повышения эффективности в принятии управленческих решений по широкому кругу вопросов. Основной целью исследования является распространение опыта анализа больших данных, которые используются в работе Министерства экономического развития Приморского края. Для этого предлагается обратиться к теоретическим аспектам больших данных, узнать основную цепочку работ над ними, а также обратить внимание на отечественный и зарубежный опыт использования данных в некоторых областях знаний. Практической частью исследования является описание опыта работы Минэкономразвития ПК, имеющее непосредственное отношение к цифровому развитию региона и осуществляющее определенный перечень работ с разнородными данными. В качестве примера обращается внимание на апробированный подход по анализу большого набора открытых данных, характеризующих работу и развитие малого и среднего предпринимательства в Приморском крае, размещенных на сайте Федеральной налоговой службы Российской Федерации. В заключении подчеркивается, что предлагаемый в исследовании подход по работе с данными может быть адаптирован под схожие потребности в других федеральных и региональных органах власти.

Ключевые слова: большие данные, анализ больших данных, цифровая трансформация, государство и большие данные, опыт работы с большими данными

Информация об авторах:

Кузора Станислав Сергеевич — ассистент Департамента инноваций Политехнического института Дальневосточного федерального университета, ORCID: 0000-0003-2309-2035, e-mail: kuzora_ss@dvfu.ru

Натаров Иван Петрович — аналитик данных Министерства экономического развития Приморского края, ORCID: 0000-0002-6550-6833, e-mail: natarov_ivan@bk.ru