DOI: https://doi.org/10.53052/9788367652018.08

Iva KOSTADINOVA[1], Vasil TOTEV[1], Aleksandra KŁOS-WITKOWSKA[2], Krzysztof WITOS[2], Marcin BERNAŚ[2], Vasyl MARTSENYUK[2], Georgi DIMITROV[1], Dejan RANCIC[3], Oleksiy BYCHKOV[4]

DOBRE PRAKTYKI W BIG DATA: ZBIERANIE SPECYFIKACJI IT

Streszczenie: Celem pracy jest zebranie i badanie specyfikacji IT dobrych praktyk w Big Data. Ankieta została przeprowadzona online za pomocą narzędzi formularzy Google. To badanie poszukuje praktycznych rozwiązań z wykorzystaniem Big Data. Ankieta zawiera łącznie 29 pytań dotyczących architektury, reprezentacji danych, przetwarzania i jakości danych, platform i narzędzi, analityki i uczenia maszynowego oraz zbiorów danych różnych projektów. Praca ta jest częścią badań w ramach IO1 w związku z celami projektu 2020-1-PL01-KA203-082197 „Innowacje dla Big Data w świecie rzeczywistym" (iBIGworld) w ramach programu Erasmus+.

Słowa kluczowe: Big Data, dobra praktyka, projekt, iBIGworld

ON GOOD PRACTICES IN BIG DATA: COLLECTING IT SPECIFICATIONS

Summary: The objective of the work is to collect and research IT specifications of good practices in Big Data. The survey was performed online using google forms tools. This research is looking for practical solutions using Big Data. The survey contains a total of 29 questions looking at Architecture, Data representation, Data processing and quality, Platforms and Tools, Analytics and Machine learning, and Data Sets of the various projects. This work is a part of the research within IO1 in connection with the objectives of project 2020-1-PL01-KA203-082197 "Innovations for Big Data in a Real World" (iBIGworld) under the Erasmus+ program.

Keywords: Big Data, good practice, project, iBIGworld

[1] University of Library Studies and Information Technologies, Sofia, Bulgaria: (i.kostadinova, v.totev, g.dimitrov)@unibit.bg
[2] Department of Computer Science and Automatics, University of Bielsko-Biala, Poland: (mbernas, kwitos, vmartsenyuk, awitkowska)@ath.bielsko.pl
[3] University of Niš, Nis, Serbia: dejan.rancic@elfak.ni.ac.rs
[4] Taras Shevchenko National University of Kyiv, Kiev, Ukraine: oleksiibychkov@knu.ua

1. Introduction

Lately a lot of attempts have been made for the purpose of implementation of Big Data solutions in different areas [1-8]. When designing innovative training courses on Big Data these good practices should be taken under consideration. In this work the research was conducted in the context of project no. 2020-1-PL01-KA203-082197 entitled “Innovations for Big Data in a Real World”. The survey was obtained by the scientist based on researching and collecting IT specifications of good practices in Big Data. The survey was performed online using google forms tools.
Due to various formats and specification of the found information in each case, the data was collected by scientists based on phrase search. Several search phrases were used: "Big Data", "good practice" and "specification". The survey was performed during a period from the 1’th of September 2020 to the 28’th of February 2021. To obtain a wide range of data multiple question fields, with an additional open-field option, were offered to mitigate the effect of narrowed answers suggestions.
The survey contains both open and closed questions. The questions consider good practices and collecting IT specifications of good practices in Big Data. To make a process of data collection unbiased no additional recommendation was added. No events were reported during that time that could influence the result.

Target

This survey is a part of the research within IO1 in connection with the objectives of project 2020-1-PL01-KA203-082197 "Innovations for Big Data in a Real World" (iBIGworld) under the Erasmus+ program. This project aims to join together Universities, business and provide innovative solutions to develop Big Data experts.
This research is looking for practical solutions using Big Data. The survey contains a total of 29 questions looking at Architecture, Data representation, Data processing and quality, Platforms and Tools, Analytics and Machine learning and Data Sets of the various projects.
The data of the research is processed by IBM SPSS Statistics 19.

2. Collection and analysis of data

The research was conducted by scientists from the 4 countries - participants in the project - Poland, Ukraine, Bulgaria Serbia. The survey contains 17 completed questionnaires for 15 found solutions, using the Big Data. IT specifications of good practices in solutions and projects, using Big Data are considered.
The survey was made without the numbering of the questions. The survey analysis process includes the title of the question, a description and an analysis of the results.
In total 17 questionaries were collected by 11 scientists - researchers.
Analysing Fig. 1, we can see that the highest number of questionnaires came from Poland - 5 (29.4%) and Bulgaria - 5 (29.4.8%), while 4 (23.5%) of questionnaires came from Serbia and 3 (17.6%) from Ukraine.

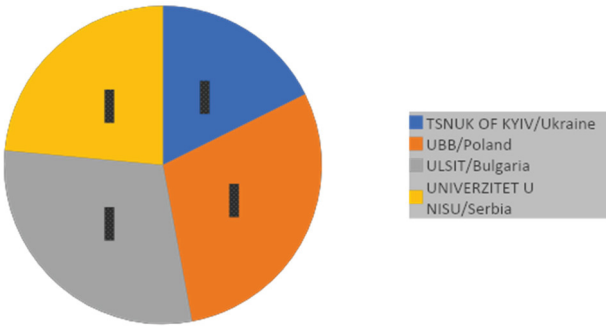

Figure 1. Distribution of surveys by nationality of scientists - researchers

3. Results on Collectiong IT Specifications of a Good Practices in Big Data

In the present research, 15 projects using Big Data were found and analyzed. Table 1 lists their titles and URL (if any).

Table 1. Titles and URL of the case/solution, in the field of Big Data

Title of the case/solution, in the field of Big Data		**URL**
1.	Netflix	*www.netflix.com*
2.	Data-Driven Bio economy	*https://www.databio.eu/en/*
3.	COVID-19 Data Hub	*https://covid19datahub.io/*
4.	WALMART How Big Data Is Used To Drive Supermarket Performance	*https://www.machinemetrics.com/blog/walmart-big-data-case-study*
5.	Air quality meter	*http://fijak-logic.com/pl/?q=node/1*
6	Model-based anomaly detection in industrial IoT systems	*No URL*
7	Big data face recognition	*https://gitlab.com/senioroman4uk/bigdata-face-recognition*
8	eDiscovery	*https://www.relativity.com/ediscovery-software/relativityone/*
9	Text (Ad) Classifier, and other projects	*https://demo.niri-ic.com/#!/app/dashboard*
10	Hearst Data Analytics Case Study	*https://aws.amazon.com/solutions/case-studies/hearst-data-analytics/*
11	Big Data in Biosensor design	*https://ibigworld.ath.edu.pl/index.php/en/*
12	Big Data In Auto Insurance & Innovative Mobility Services	*https://trackandknowproject.eu/*
13	Big Data Innovations In Fleet Management – Vodaphone Innovus	*https://trackandknowproject.eu/*
14	D2Lab - Data Diagnostic Laboratory	*https://d2lab.nissatech.com/*
15	Big data image recognition	*No URL*

The projects and solutions listed in this way are considered according to a number of indicators. In the following points, we will show and analyze the status of each solution/project according to these criteria.

3.1. Country - the origin of the case/solution in the field of Big Data

The first question concerns in which country the case/solution was implemented. There is no restriction on the search for such a case/solution. The research is global, in the whole world.
The cases/solutions in the field of Big Data can be created in one country, but they can also be created under an international project (scientific, European, Erasmus), in which many countries (from the EU or not from the EU) take part. In case they are realized under a project - then the country of the leading organization is marked as a country.

Data description (Table 2)

Table 2. Country - the origin of the case/solution in the field of Big Data

Country - the origin of the case/solution in the field of Big Data		**Frequency**	**Percent**
Valid	Belgium	1	6,7%
	Canada	1	6,7%
	Greece	1	6,7%
	Italy	1	6,7
	Poland	2	13,3
	Serbia	2	13,3
	Ukraine	3	20,0
	United Kingdom	1	6,7
	USA	3	20,0
	Total	15	100,0

Discussion

Researchers was found the most case/solution with country of origin USA - 3 projects and Ukraine - again 3 projects. The following projects are from Poland and Serbia - 2 projects for each of these 2 countries. Researchers have described 1 project each from Belgium, Canada, Greece, Italy and the United Kingdom.
The considered cases/solutions can be grouped according to the criterion of whether their country is a member of the EU or not. The data for this are given in Table 3.

Table 3. EU/No EU countries with the case/solution in the field of Big Data

EU/No EU countries with the case/solution in the field of Big Data		**Frequency**	**Percent**
Valid	EU countries	5	33,3%
	No EU countries	10	66,7%
	Total	**15**	**100,0**

The found and described cases/solution in the field of Big Data come from 10 non-EU countries, and only 5 are EU members. Figure 2 shows graphically the distribution

of the found solutions according to whether the organizing country is a member of the EU or not.

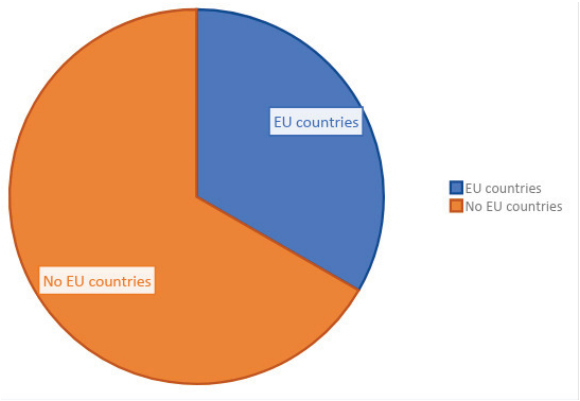

Figure 2. Distribution of the found solutions according to whether the organizing country is a member of the EU or not

Considering the described cases and countries in which they are implemented, we can make cross following analysis (Table 4):

*Table 4. Country - the origin of the case/solution in the field of Big Data * EU/No EU countries Crosstabulation*

Country - the origin of the case/solution in the field of Big Data	**EU/No EU countries**		**Total**
	EU	**No EU**	
Belgium	1	0	1
Canada	0	1	1
Greece	1	0	1
Italy	1	0	1
Poland	2	0	2
Serbia	0	2	2
Ukraine	0	3	3
United Kingdom	0	1	1
USA	0	3	3
Total:	**5**	**10**	**15**

Discussion

Most of the described projects in the field of Big Data are from the USA and Ukraine - 3 in number for each of these countries. Both countries are not members of the EU. Other projects from non-EU countries are Serbia with 2 projects, the United Kingdom with 1 solution and Canada with 1 project.

Poland has the most projects like an EU member state - 2 projects, followed by Italy and Greece with 1 project each. In general - the countries participating in this project have considered their own case/solution and solutions of leading countries in the IT industry.

3.2. Does company is using solutions based on open-source?

It is investigated whether the companies to the researched projects use solutions based on open source. Table 4 contains the data on this indicator.

Table 4. Does company is using solutions based on open-source?

Does company is using solutions based on open-source?		Frequency	Percent
Valid	No, only proprietary	2	13,3
	Only open source	6	40,0
	Partially	7	46,7
	Total	**15**	**100,0**

It can be seen that the majority of organizations use sources that are partially open-source - 7 of the described projects or 46.7% of the surveyed. Quite a large number of organizations use only open-source solutions - 6 (40%), while only 2 (13.3%) of the organizations use entirely their own sources.

3.3. Does company is using open-source data sources?

It is researched whether the organizations use open-source data sources. Summary data on the extent to which organizations use open-source data sources are shown in Figure 3, and detailed information on the project name and its characteristics are shown in Table 5.

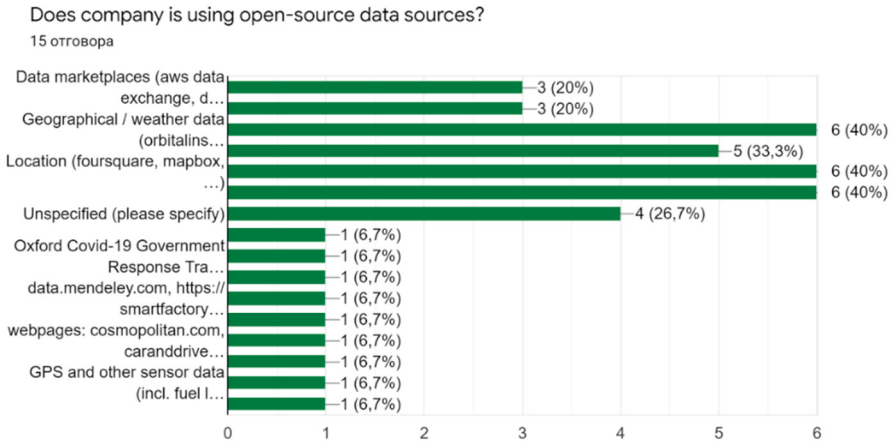

Figure 3. Distribution by the what kind of open-source data sources are used in the project.

Table 5. Does company is using solutions based on open-source (by the company)

Company*	**Data marketplaces (aws data exchange, dawex)**	**Financial & economic data (Bloomberg, Thomson Reuters, …)**	**Geographical / weather data (orbitalinsights, windward, Airobotics, …)**	**People/ Entities (zoominfo, acxiom, …)**	**Location (foursquare, mapbox, …)**	**Other (data.gov, IMAGEnet, …)**	**Unspecified (please specify)**
1.Netflix				Yes	Yes		
2.Data-Driven Bio economy			Yes		Yes		telemetry data, IoT, NoSQL, Media Image, SQL, Graph metadata, BI
3. COVID-19 Data Hub			Yes			Yes	Oxford Covid-19 Government Response Tracker, *https://www.google.com/covid19/mobility/, https://www.apple.com/covid19/mobility/*
4. Walmart How Big Data Is Used To Drive Supermarket Performance	Yes	Yes	Yes	Yes	Yes		
5. Air quality meter					Yes		*https://www.openstreetmap.org*
6. Model-based anomaly detection in industrial IoT systems						Yes	*data.mendeley.com, https://smartfactory-owl.de/, kaggle.com*
7. Big data face recognition		Yes		Yes	Yes		
8. eDiscovery						Yes	
9. Text (Ad) Classifier, and other projects			Yes				News articles, Images
10. Hearst Data Analytics Case Study	Yes						webpages: *cosmopolitan.com, caranddriver.com*
11. Big Data in Biosensor design							Yes

12. Big Data In Auto Insurance & Innovative Mobility Services			Yes		Yes		GPS location data from vehicle black boxes, historic telematics, environmental, demographic and geographic information
13. Big Data Innovations In Fleet Management – Vodaphone Innovus			Yes	Yes	Yes		GPS and other sensor data (incl. fuel level and driver behavior data) from vehicles (tracks) and their drivers
14. D2Lab - Data Diagnostic Laboratory							Manufacturing data, real-time data streams from many machines and industrial processes (Industry 4.0)
15. Big data image recognition		Yes		Yes		Yes	

Discussion

From the data in Figure 3 and Table 5, it can be summarized that companies use mainly geographically open-source data sources and those that collect location data. But there are also many other different solutions based on open-source.

3.4. Does company is using open-source resources?

Table 6 shows what open-source resource used in companies for their projects. Different possibilities for open-source resources are reflected. Like a Web mining, different researching's from an organization like OpenAI or Vector Institute, different data services, and Connectors and API's to all major data historians vendors.

Table 6. Does company is using open-source resources

Does company is using open-source resources?		**Frequency**	**Percent**
Valid	Connectors and API's to all major data historians vendors such as OSIsoft and Honeywell, and output results into our user interface or integrate with an existing solution	1	6,7%
	Data services (quantum black, Kaggle, ElectrifAI)	4	26,7%
	No answer	3	20,0%
	Research (OpenAI, Vector Instutute, …)	6	40,0%
	Web mining	1	6,7%
	Total	**15**	**100,0**

As many as 6 (40%) of the projects described in the study use research companies specialized in the field of artificial intelligence. Another 4 (26.7%) use Data services (like Quantum black, Kaggle, ElectrifAI) and only one (6,7%) use web mining. For three of the examined projects, no information was found whether they use open-source resources. Figure 4 graphically depicts this data from Table 6.

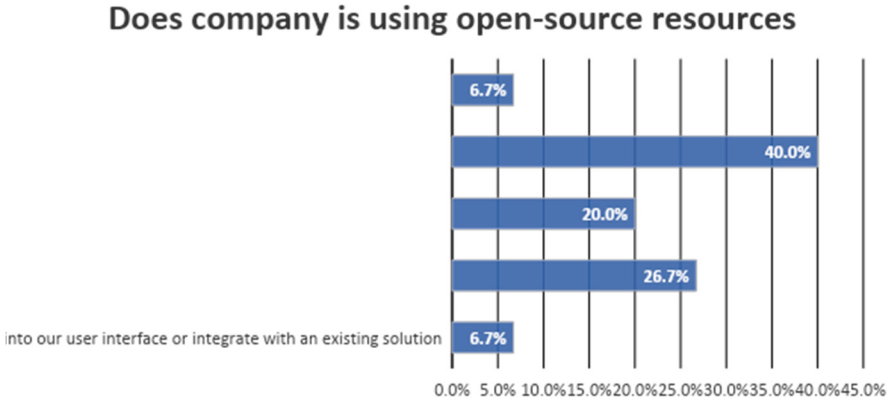

Figure 4. Distribution by the kind of using open-source resources

3.5. What is the result of processing the data? What kind of information is retrieved?

The data for area of implementation of the Big Data solution and What kind of information is retrieved of processing the data are reviewed in the Table 7.

Table 7. Area of implementation of the researched projects

What is the area of implementation of the Big Data solution?		**Number**	**Percent**	**What is the result of processing the data**
Valid	Agriculture	1	6,7%	maps for variable rate application, solution in form of web GIS
	Business cases: Insurance, Electric cars, Car pooling	1	6,7%	Insurance: in-depth and accurate crash probability estimation, Electric cars: (i) cost-benefit of a switching to an electric car mobility; (ii) matching global charging times and charging points to drivers' habits, Car Pooling: (i) park decreasing due to sharable routes; (ii) cost- benefit of switching to a sharing mobility paradigm; (iii) likelihood of finding a proper sharable route that matches time and geographical zone
	C2C businesses	1	6,7%	Semantic relatedness of phrases, Classification models, Input correction & completion models
	Data-driven analytics solutions for manufacturing, transportation, additive manufacturing and oil & gas firms.	1	6,7%	Detect unusual service behaviour. The system ensures early and precise detection of unusual behaviour in large problem/process spaces. Prevent asset failure, detect quality issues and improve operational processes of customer business. Enable highly personalized service offering to an individual customer.

				D2Lab solution increases efficiency of a supply chain by predicting demand and reducing wasteful stockpiling. It can spot anomalies in logistic process and improve it accordingly.
	Entertainment services (tv, movies, shows)	1	6,7%	What titles customers watch, what time of day movies are watched, time spent selecting movies, how often playback is stopped, delays caused by buffering, bitrate, customer location
	Fleet Management	1	6,7%	•Predictive maintenance • Anomaly detection, reduction of false alarms • Correlation of Fleet Data with external Weather and Traffic services • Fleet costs reduction • Fleet downtime reduction • Fleet response time improvement • Improve driver behavior and reduce accidents
	Healthcare	1	6,7%	providing the research community with a unified dataset by collecting worldwide fine-grained case data, merged with exogenous variables helpful for a better understanding of COVID-19
	Industrial automation and diagnosis	1	6,7%	Identification of anomalous behavior, malfunctioning or wear, an root causes in the system
	Information security, marketing, decision support ,banking	1	6,7%	Customer decision support and recomendation, security alerts, analysis results
	Law firms	1	6,7%	Discovery in legal proceedings such as litigation, government investigations, or Freedom of Information Act requests.
	Marketing, solutions upholder, data defence	1	6,7%	Analysis of bissnes-processes
	media and information	1	6,7%	aggregated data—available to editors in minutes
	Retail and Wholesale trade	1	6,7%	
	The data could be used in biosensor design as a part of cyber biophysical system	1	6,7%	The obtain data concerning parameters of functioning biosensor devices, their operational and self stability.
	Weather data processing	1	6,7%	Information on air quality in large urban areas
	Total	**15**	**100,0**	

Discusion

The projects described in the study are in different areas. There is no overlap in the area of implementation of the Big Data solution. Each of the projects has a different application and generated different kind of information.

3.6. What kind of applications/tools are used in data Insight/Consume stage (e.g. Tableau, R Studio, other)?

The question "What kind of applications / tools are used in data Insight / Consume stage" received many answers. In most cases, more than one program is used in different companies. However, there are some applications that are used more often. The data for them are placed in table 8 and are visualized in Figure 5.

Table 8. What kind of applications/tools are used in data Insight/Consume stage

What kind of applications/tools are used in data Insight/Consume stage?		**Frequency**	**Percent**
Valid	Python	3	20,0%
	R Studio	3	20,0%
	Jupyter	5	33,3%
	matlab	1	6,7,0%
	Amazon Web Services	1	13,3%
	Others	2	6,7%
	Total	15	100,0

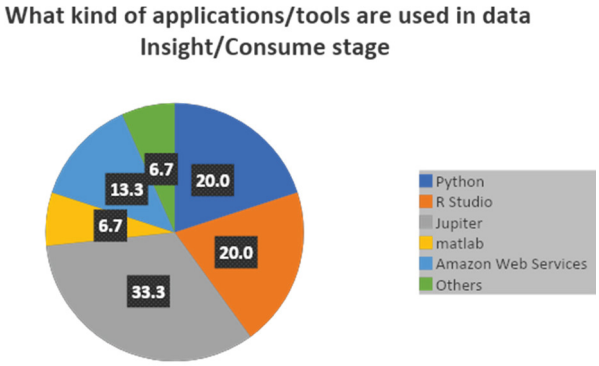

Figure 5. Distribution by the kind of applications/tools are used

Discussion

The data show that Jupiter is the most used - in 5 out of 15 projects, followed by Python and R studio - in 3 out of 15 studied projects.

3.7. What type of licensing is used for the solution?

Regarding the question “What type of licensing is used for the solution” - 4 (or 27%) of the projects use Proprietary license, other4 (or 27%) of the projects use Copyleft (GPL, LGPL) license. Two of the projects use Permissive licensing (13%) and the other 2 projects (13%) use Open licensing. The data are shown in Figure 6.

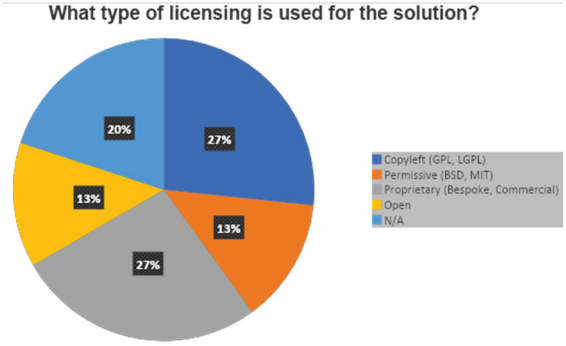

Figure 6. Distribution by type of licensing is used for the solution

3.8. Describe the architecture of the Big Data solution including the process scheme using the following steps

Big data solutions typically involve one or more of the following types of workload. These are ingest, store, transform, analyse and insight / application and designing tools.

It is interesting to see the architecture of the studied Big Data solutions. Are these steps of workloads also is implemented in the solutions, described in this research?

Figure 7 describes a sequence of processes comprising several steps. It is reflected which of these processes are realized in the processes participating in the study. The data are reflected graphically.

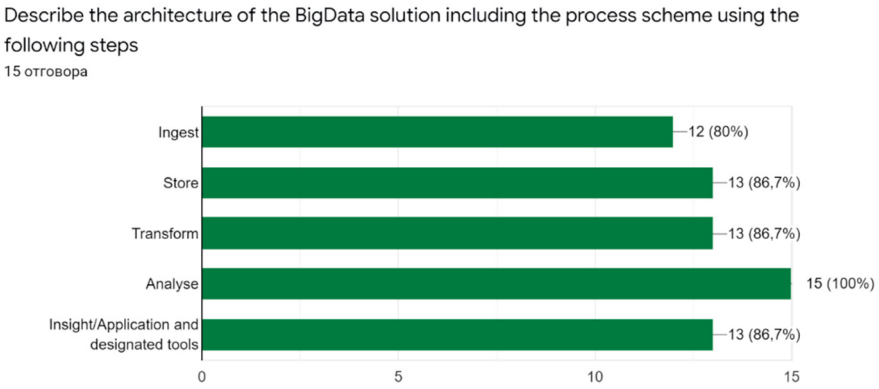

Figure 7. Description the architecture of the Big Data solution

The data in Figure 7 show that almost all of the listed typical steps are present in almost all of them - are ingest, store, transform, analyse and insight/application and designing tools. All 15 projects include data annals.

Only 13 of them, however, do not include ingest, store and insight/application and designed tools. Their activity is mainly limited to data analysis.

Ingest is not used in three of the solutions described. This step is necessary if the solution includes real-time sources, and the architecture must include a way to capture and store real-time messages for stream processing. This means that 3 of the solutions described in this way do not need such a buffer.

3.9. What is the source of data

Attention is paid to what is the source of data in the studied projects. Followed is whether the data comes from a database or collected by any service, application, sensor or Web (Fig.8).

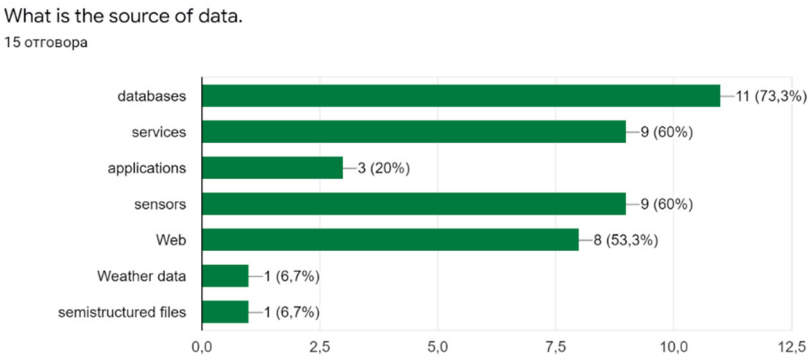

Figure 8. Description by the source of data

The projects use several sources for data accumulation. Data contained in databases are most often collected - in as many as 11 of the projects or 73.3%. To a lesser extent, but still, strongly enough, these Big Data solutions collect data from services and sensors - in 9 (60%) of the studied projects. Web data is collected in 8 of these projects (ie 53.3% of the projects). Only 20% of projects collect their data from the use of applications, and only 6.7% - from weather data or semi-structured files.

3.10. What is the volume of data process?

Figure 9 shows what is the volume of data processing in researchers Big Data solutions.

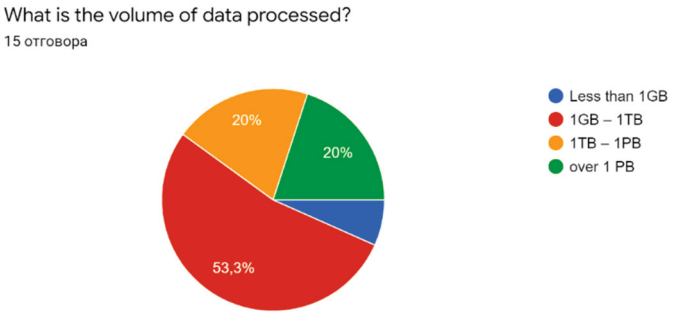

Figure 9. Description by the volume of data process

In 53.3% of solutions, the volume of data processing is between 1GB and 1TB. In 20% of projects the volume of data is between 1TB and 1PB. Also in 20% the

volume of data is over 1PB. In a very small part of the projects, the information is less than 1GB.

3.11. What is the data characteristics?

Figure 10 shows the distribution regarding the characteristics of the data. The largest share is represented by data in the form of records in noSQL databases - 33.3%, followed by records in SQL databases by 20%. The remaining 46.7 are distributed among many other types of data - files (data, picture, sound, video), key-value pairs, graphs, time, series, csv files, JSON files, TDMS file, RDBMS.

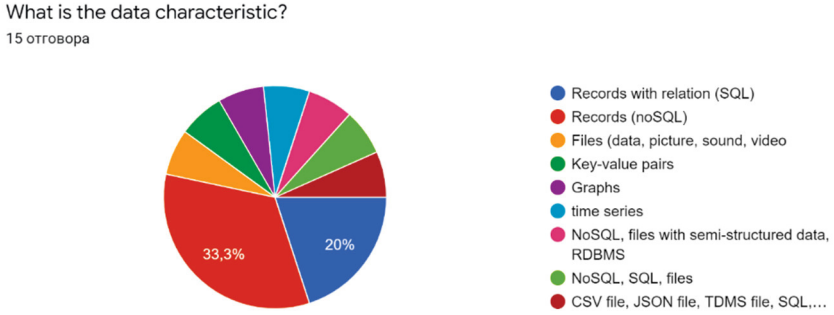

Figure 10. Description by the data characteristics

Table 9 provides a cross-analysis between the different data types and their volume. It is noteworthy that most of them are records in SQL and noSQL data bases, and their size is over 1GB.

*Table 9. Cross analyses: What is the data characteristic? * What is the volume of data processed?*

		What is the volume of data processed?				Total
		1GB – 1TB	1TB – 1PB	Less than 1GB	over 1 PB	
What is the data characteristic?	CSV file, JSON file, TDMS file, SQL, NoSQL	1	0	0	0	1
	Files (data, picture, sound, video	0	0	1	0	1
	Graphs	1	0	0	0	1
	Key-value pairs	0	1	0	0	1
	NoSQL, files with semi-structured data, RDBMS	1	0	0	0	1
	NoSQL, SQL, files	1	0	0	0	1
	Records (noSQL)	2	0	0	3	5
	Records with relation (SQL)	2	1	0	0	3
	time series	0	1	0	0	1
Total		8	3	1	3	15

2.12. What is the data tier? (e.g. records [SQL, noSQL based], files [types],key-value pairs, graphs or others)

Table 10 gives an idea of how the data in the considered projects are organized. Each of the projects has found a specific way to organize their data.

Table 10. Description by the data tier?

What is the data tier? (e.g. records [SQL, noSQL based], fi_x000C_les [types],key-value pairs, graphs or others)		Frequency	Percent
Valid	csv files, stored dataframes	1	6,7
	customer data, noSQL based	1	6,7
	key-value	2	13,3
	key-value as clickstream data	1	6,7
	NoSQL	1	6,7
	NoSQL, HDFS	1	6,7
	NoSQL, SQL based	1	6,7
	records	1	6,7
	records, graphs	1	6,7
	SQL	2	13,3
	SQL, NoSQL (HBASE)	1	6,7
	telemetry data, IoT, NoSQL, Media Image, SQL, Graph metadata, BI	1	6,7
	XML exports of relational databases	1	6,7
	Total	15	100,0

3.13. What tools are used to store data (NoSQL, NewSQL, Graph databases, Server-less, Cluster SVCS, others?

Table 11 describes the tools are used to store data. Among them, NoSQL is most often used - in 3 of the analyzed projects.

Table 11. Description by the tools are used to store data

What tools are used to store data		Frequency	Percent
Valid	Apache Hive, Apache HBase, MongoDB	1	6,7
	Cassandra, HBase	1	6,7
	Cassandra, HBase, HDFS	1	6,7
	Collections of files	1	6,7
	Git LFS	1	6,7
	Graph databases	1	6,7
	Key-value databas (DynamoDB), AWS S3	1	6,7
	MongoDB, HBase, HDFS	1	6,7
	MySQL	1	6,7
	NoSQL	3	20,0
	RelativityOne	1	6,7
	Server-less	1	6,7
	Telemetry data, IoT, NoSQL, Media Image, SQL, Graph metadata, BI	1	6,7
	Total	15	100,0

3.14. What is a velocity of data?

Data velocity is the speed at which data is processed. This includes input such as processing of social media posts and output such as processing required to produce a report or execute a process. In this study, attention is drawn what is the velocity of data in the surveyed 15 projects.
Figure 11 shows that data velocity is both hot and cold for as many as 13 (86.7%) of projects. One of the projects has hot data, and there is no information about another.

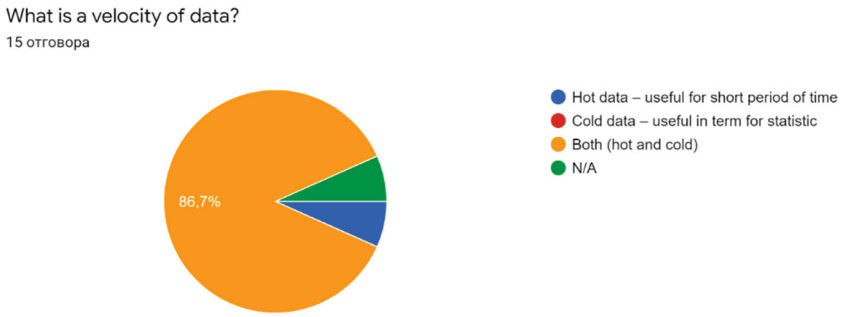

Figure 11. Description by the data characteristics

3.15. Describe the Variety of Big Data (structure of data and initial relations (one type of data/ multiple - types)

Table 12 describes the Variety of Big Data for each of the projects in the study. There is a rich variety for all 15 project.

Table 12. Describtion by the variety of Big Data (structure of data and initial relations (one type of data/ multiple - types)

What tools are used to store data		**Frequency**	**Percent**
Valid	CSV file, JSON file, TDMS file, SQL, NoSQL	1	6,7
	customerID, movie ID, rating, date the movie was watched, time spent selecting movies, how often playback is stopped,delays caused by buffering, bitrate, customer location	1	6,7
	GPS and other sensor data (incl. fuel level and driver behavior data) from vehicles (tracks) and their drivers enriched with contextual data: weather information, POIs,...	1	6,7
	Hierarchical data, Graph data, Lists... all sorts of data used.	1	6,7
	multiple data types	1	6,7
	multiple types of data	1	6,7
	multiple types of data; namely, 1) standard COVID-19 variables: total population, cumulative number of cases, tests, deaths, recovered, daily number of hospitalized, patients requiring ventilation and intensive therapy; 2) policy measures by Oxford COVID-19 Government Response Tracker; 3) geographic information suited for	1	6,7

	data visualization and for interfacing with external databases; 4) external identifiers allowing to extend the dataset with World Bank Open Data, Google mobility reports, and Apple mobility reports.		
	multiple-types	1	6,7
	N/A	2	13,3
	one type - web events	1	6,7
	Raw GPS points semantically enriched with external information, e.g. attaching to each GPS location the weather conditions, local traffic and points-of-interest around it	1	6,7
	Records: Air quality, data and time, location	1	6,7
	structured data in tables, picture data, text	1	6,7
	time series of multidimensional analog signals	1	6,7
	CSV file, JSON file, TDMS file, SQL, NoSQL	1	6,7
	customerID, movie ID, rating, date the movie was watched, time spent selecting movies, how often playback is stopped,delays caused by buffering, bitrate, customer location	1	6,7
	Total	15	100,0

3.16. What a veracity (quality) of data?

Regarding veracity or data quality - 80% of them are data with noise. Only 13.3% is only vital information, and for the remaining 6.7% there is no information. The data are shown in Fig.12.

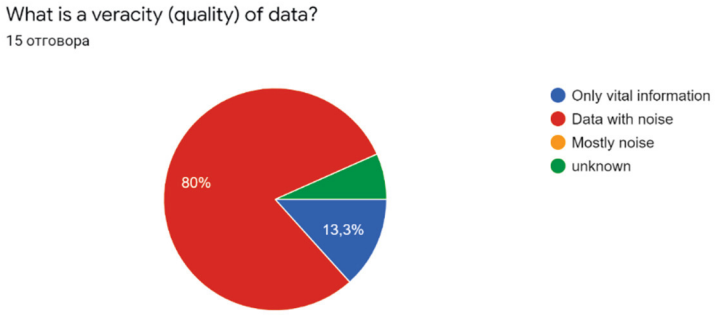

Figure 12. Description by veracity (quality) of data

3.17. Are security features is supported in projects?

Figure 13 shows whether the studied projects have security features. The majority of projects - 7 (46.7%) support such functions. Another 5 (33.3%) have no security features, and the remaining 3 (20%) have only partial ones.

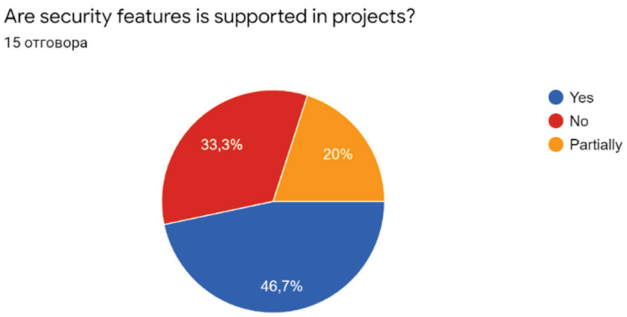

Figure 13. Description by veracity (quality) of data

3.18. What tools are used when processing data and to ensure their quality?

Table 13 lists the various tools that were used in the data processing process to ensure its quality. Most projects use different tools. The NoSQL tool takes the biggest part for this purpose - in 3 of the researched projects.

Table 13. Description by the tools are used when processing data and to ensure their quality

What tools are used to store data (NoSQL, NewSQL, Graph databases, Server-less, Cluster SVCS, others? Please provide the tools names		**Frequency**	**Percent**
Valid	Apache Hive, Apache HBase, MongoDB	1	6,7
	Cassandra, HBase	1	6,7
	Cassandra, HBase, HDFS	1	6,7
	Collections of files	1	6,7
	Git LFS	1	6,7
	graph databases	1	6,7
	key-value database (DynamoDB), AWS S3	1	6,7
	MongoDB, HBase, HDFS	1	6,7
	MySQL	1	6,7
	NoSQL	3	20,0
	RelativityOne	1	6,7
	server-less	1	6,7
	telemetry data, IoT, NoSQL, Media Image, SQL, Graph metadata, BI	1	6,7
	Total	15	100,0

3.19. What Big Data platform type is used (e.g. server based, cloud solutions, with/without edge computing support or other)?

Regarding the type of platform used for Big Data - the considered projects are implemented on 2 types of platforms - dedicated server and cloud solutions. 14 (93.3%) of the projects use cloud solutions, and 5 (33.3%) use a dedicated server (Fig. 14).

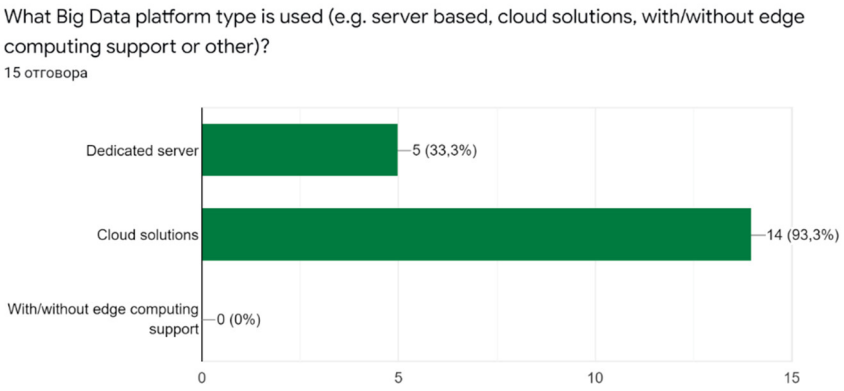

Figure 14. Description by Big Data platform type is used

However, in a cross-analysis, it can be seen that some projects use both types of Big Data platform. The data are presented in Table 14.

*Table 14.Cross analyses Title of the solution * What Big Data platform type is used?*

		What Big Data platform type is used			
		Cloud solutions	Dedicated server	Dedicated server, Cloud solutions	Total
Title of the case/ solution	Air quality meter	0	1	0	1
	Big data face recognition	1	0	0	1
	Big data image recognition	1	0	0	1
	Big Data In Auto Insurance & Innovative Mobility Services	0	0	1	1
	Big Data in Biosensor design	1	0	0	1
	Big Data Innovations In Fleet Management - – Vodaphone Innovus	0	0	1	1
	COVID-19 Data Hub	1	0	0	1
	D2Lab - Data Diagnostic Laboratory	0	0	1	1
	eDiscovery	1	0	0	1
	Hearst Data Analytics Case Study	1	0	0	1
	Horizon 2020 - Data-Driven Bioeconomy - Data Bio Project ;	1	0	0	1
	Model-based anomaly detection in industrial IoT systems	1	0	0	1
	Netflix	1	0	0	1
	Text (Ad) Classifier, and other projects	0	0	1	1

	WALMART How Big Data Is Used To Drive Supermarket Performance	1	0	0	1
Total		10	1	4	15

3.20. What platform solution is used?

Figure 15 describes the platforms that were used in the studied projects. Most of the projects use the MongoDB - 4 platform (26.7%). Next are the projects using Microsoft Azure - 3 (20.0%). They are followed by Apache Hadoop / HDFS / HBase, Spark, Kafka, Kafka treams and Tableau, used in 2 projects (Fig. 15).

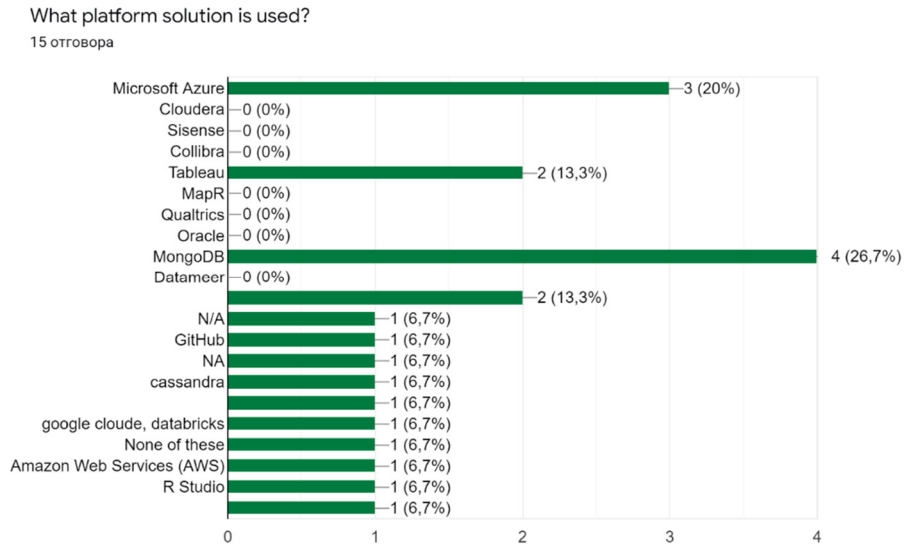

Figure 15. Description by platform solution is used

3.21. What kind of tools/application is used for data extraction/ingestion (e.g. Apache Kafka, script - node.js, others)?

Various applications have been used to extract data from these projects. In some projects only one application was used, in others several were used. Table 15 contains what applications (one or in combination) are used in the different projects.

Table 15. Description by tools/application is used for data extraction/ingestion ?

What tools are used to store data (NoSQL, NewSQL, Graph databases, Server-less, Cluster SVCS, others? Please provide the tools names		**Frequency**	**Percent**
Valid	Apache Kafka	5	33,3
	Apache Kafka, script - python, node.js, visualisation - web application	1	6,7
	N/A	2	13,3
	node.js	1	6,7

	Python	2	13,3
	Python & Jupyter, Apache Kafka	1	6,7
	R Studio	2	13,3
	script - JavaSctipt, C#	1	6,7
	Total	15	100,0

Discusion

The most common data retrieval tool is the Apache Kafka software platform - it is preferred in 5 (33.3%) of the projects as a standalone product and in 2 other projects in combination with other tools. In second place after it is R Studio, which is used in 2 (13.3%) of the projects as a standalone product. Other tools used are Java script and node.js, which are used alone in a project. For 2 of the projects there is no information what tool they use.

3.22. What type of storage is used?

The research also traced what type of data storage was used in the described projects. Figure 16 shows the distribution regarding the storage location of the project information.

Nine of the projects (60%) use a cluster to store information. Three (20%) of the projects use Stream based, and the remaining projects use Data lake, Data hub or File system and Relational DataBase.

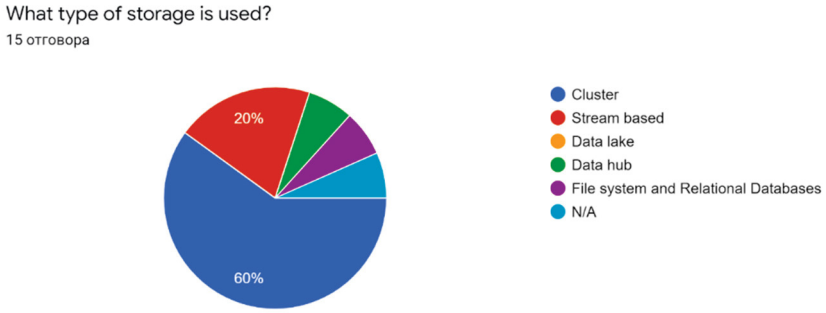

Figure 16. Description by type of storage used

3.23. What kind of additional applications/tools are used in data storage?

Table 16 describes additional tools and applications for data storage and management used in the studied projects. HADOOP is used in 7 (46.7%) of these projects. They are followed by storage in cloud platforms such as Amazon S3, Azure Storage, Google cloud storage - in 4 (26.7%) of the projects.

Table 16. Description by additional applications/tools are used in data storage

What kind of additional applications/tools are used in data storage		Frequency	Percent
Valid	Data Warehouses (Amazon redshift, Google Big Querry, …)	1	6,7
	HADOOP (Cloudera, Amazon EMR, Google Cloud dataprc, …)	7	46,7
	N/A	2	13,3
	Storage (Amazon S3, Azure Storage, Google cloud storage, …)	4	26,7
	Streaming/In memory (Giga spaces, SAP cloud platform, …)	1	6,7
	Total	15	100,0

The data from the Table 16 are presented graphically in Figure 17.

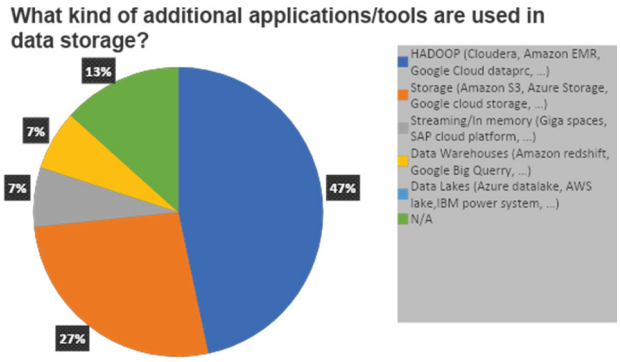

Figure 17. Description by of additional applications/tools are used in data storage

3.24. What kind of analytics is performed?

Figure 18 shows how the data generated by the projects is processed and what type of analysis is performed. In 9 (60%) of the projects classical machine learning is used. In 4 (26.7%) of the projects Deep learning is performed, and in the remaining 2 (13.3%) - statistical processing.

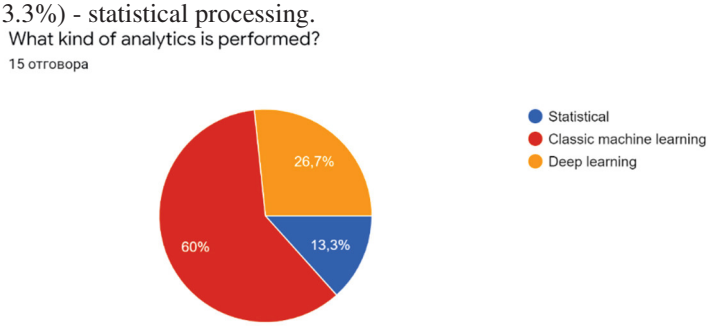

Figure 18. Description by of additional applications/tools are used in data storage

3.25. What kind of applications/tools are used in analytics stage? (dedicated to a data type / general)?

In the analysis stage, each of the described projects uses a different tool. The data for this are shown in Table 17 and visualized in Figure 18.

Table 17. Description by used applications/tools in analytics stage

What kind of applications/tools are used in analytics stage		Frequency	Percent
Valid	BI platforms (looker, amazon QuickSight, …)	1	6,7
	Data Analyst platforms (Microsoft, pentaho, alteryx, …)	1	6,7
	Data science notebooks (binder, colab, saturnCloud, …)	1	6,7
	Data Science Platforms (databricks, knime, matlab, …)	6	40,0
	Machine learning (Azure Machine learning, DataRobot, …)	3	20,0
	N/A	1	6,7
	Visualisation (tableau, Power BI, Google data studio,…)	1	6,7
	Web/ mobile analytics (google analytics, mixpanel, …)	1	6,7
	Total	15	100,0

Discussion

The preferred tool for information processing are Data Science Platforms in 6 (40%) of the projects. It is followed by Machine learning through Azure Machine learning, DataRobot and others, preferred in 3 (20%) of the projects. Tools such as the use of BI platforms, Data Analyst platforms, Data science notebooks, Visualisation or Web/ mobile analytics are used only in one of the projects.

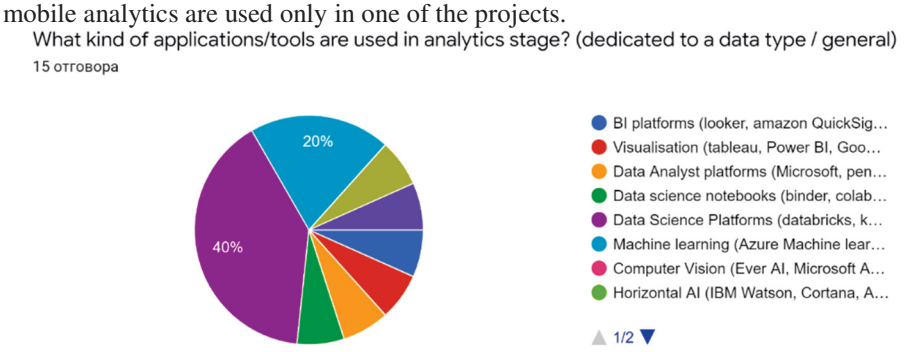

Figure 19. Description by used applications/tools in analytics stage

3.26. What programming languages are used?

The preferred language for use in such projects is Python - it is used in 10 (66.7%) of them. In second place is Java - in 7 (46.7%) of the project.
In third place is R (40%). Use of other programming languages such as c #, C ++, Matlab, JavaScript, Scala are very small. The data are shown in Figure 20.

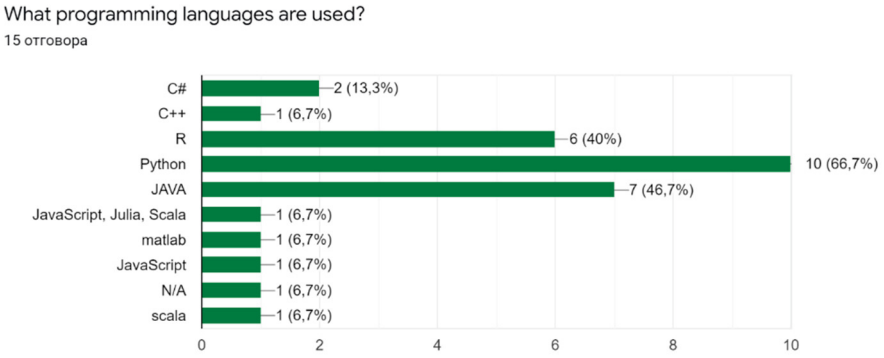

Figure 20. Description by of additional applications/tools are used in data storage

3.27. What software is used for Analytics and Machine learning?

Regarding the software used for analysis and machine learning, in some of the projects only one product was used, while in others several were used. The most commonly used is a combination of several products - in 3 (20%) of the projects - Apache Spark, Python scikit-learn. In 2 (13.3%) of the projects selected only Apache Spark of machine learning and other 2 (13.3%) is selected only R Studio.
Data on used software for Analytics and Machine learning are mentioned in Table 18.

Table 18. Description by used software for Analytics and Machine learning

What software is used for Analytics and Machine learning		Frequency	Percent
Valid	Apache Spark	2	13,3
	Apache Spark, Python scikit-learn	3	20,0
	Apache Spark, ScikitLearn, TensorFlow	1	6,7
	AWS lambda	1	6,7
	N/A	1	6,7
	NumPy, Scikit-learn, TensorFlow	1	6,7
	R Studio	2	13,3
	R Studio and SAS	1	6,7
	Relativity One	1	6,7
	Scikit learn, Keras, Matlab ml	1	6,7
	use our own algorithms, heat map	1	6,7
	Total	15	100,0

3.28. What type of data set is used in the solution?

The research noted what type of data set is used in the solutions/cases. The data are placed in a figure 21. Goals 10 (66.6%) use free date sets, while others 5(33.3%) - used paid.

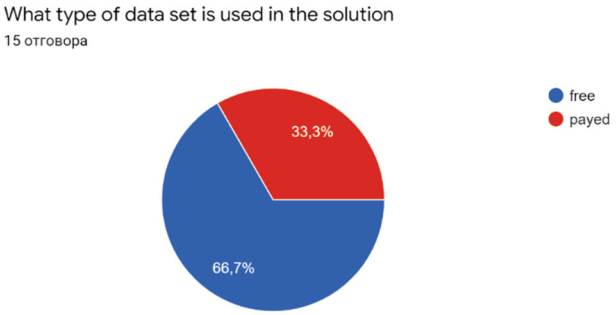

Figure 21. Description by type of data set is used in the solution

4. Conclusions

The objective of the work was to collect and research IT specifications of good practices in Big Data. The survey was performed online using google forms tools. This research was looking for practical solutions using Big Data. The survey contains the questions looking at Architecture, Data representation, Data processing and quality, Platforms and Tools, Analytics and Machine learning, and Data Sets of the various projects.

REFERENCES

1. BELLANDI V.: A Big Data Infrastructure in Support of Healthy and Independent Living: A Real Case Application (2023) Intelligent Systems Reference Library, 229, pp. 95-134. DOI: 10.1007/978-3-031-11170-9_5.
2. ARIFIN S., SILALAHI F.E.S., PRAYITNO M., MAJID N.K., AMHAR F., GULARSO H.: Geospatial Big Data Management Testing Using Open Source Technology (2023) Mechanisms and Machine Science, 121, pp. 29-42. DOI: 10.1007/978-3-031-09909-0_3.
3. ZHANG S., OU W., REN G., WANG H., ZHU P., ZHANG W.: Risk Model and Decision Support System of State Grid Operation Management Based on Big Data (2023) Lecture Notes on Data Engineering and Communications Technologies, 122, pp. 419-427. DOI: 10.1007/978-981-19-3632-6_51.
4. JOSE D.T., HOLME J., CHAKRAVORTY A., RONG C.: Integrating Big Data and blockchain to manage energy smart grids—TOTEM framework (2022) Blockchain: Research and Applications, 3 (3), art. no. 100081, . DOI: 10.1016/j.bcra.2022.100081.
5. ESPOSITO S., ORLANDI S., MAGNACCA S., DE CURTIS A., GIALLUISI A., IACOVIELLO L.: on behalf of The Neuromed Clinical Network Big Data and Personalised Health Investigators, Clinical Network for Big Data and Personalized Health: Study Protocol and Preliminary Results (2022) International Journal of Environmental Research and Public Health, 19 (11), art. no. 6365. DOI: 10.3390/ijerph19116365.

6. ZHANG G. PYCLKDE: A Big Data-enabled high-performance computational framework for species habitat suitability modeling and mapping (2022) Transactions in GIS, 26 (4), pp. 1754-1774. DOI: 10.1111/tgis.12901.
7. DEEPA N., PHAM Q.-V., NGUYEN D.C., BHATTACHARYA S., PRABADEVI B., GADEKALLU T.R., MADDIKUNTA P.K.R., FANG F., PATHIRANA P.N.: A survey on blockchain for Big Data: Approaches, opportunities, and future directions (2022) Future Generation Computer Systems, 131, pp. 209-226. DOI: 10.1016/j.future.2022.01.017.
8. KASTOUNI M.Z., AIT LAHCEN: A. Big data analytics in telecommunications: Governance, architecture and use cases (2022) Journal of King Saud University - Computer and Information Sciences, 34 (6), pp. 2758-2770. DOI: 10.1016/j.jksuci.2020.11.024.