DOI: https://doi.org/10.53052/9788367652018.01

Marcin BERNAŚ[1], Iva KOSTADINOVA[2], Vasil TOTEV[2],
Vasyl MARTSENYUK[1], Georgi DIMITROV[2], Dejan RANCIC[3],
Oleksiy BYCHKOV[4]

O METODOLOGII A1.1 ZBIERANIA DOBRYCH PRAKTYK BIG DATA: PROJEKTOWANIE BADAŃ I WNIOSKOWANIE A1.2

Streszczenie: Praca poświęcona jest opracowaniu metodologii zbierania dobrych praktyk z zakresu Big Data. Następnie podana technika została zastosowana do projektowania badań i analizy wyników. Szczególną uwagę przywiązujemy do umiejętności twardych i miękkich oraz tematów, które powinny być poruszane na szkoleniu Big Data. Niniejsza praca jest częścią badań w ramach IO1 w związku z celami projektu 2020-1-PL01-KA203-082197 „Innowacje dla Big Data in a Real World" (iBIGworld) w ramach programu Erasmus+.

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

ON METHODOLOGY A1.1 FOR COLLECTING BIG DATA GOOD PRACTICE: DESIGNING RESEARCH AND CONCLUDING A1.2

Summary: The work is devoted to the development of methodology for collecting good practices in the field of Big Data. Then given technique has been applied to designing the research and analyzing the results. Particular attention is paid to the hard and soft skills as well as the topics that should be covered by the Big Data training course. 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, innovation, iBIGworld

Introduction

The project iBIGworld preassumes the fundamental study of existing practice in the field of Big Data. That is why we have started the project from deep

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

and comprehensive research of good practices. The objective of the given work is to get the list of the most needed skills in Big Data. We start from the development of the methodology of the collection of the Big Data good practices and implementing in the form of the survey study. Further, in order to evidence the corresponding reports, the results of the surveys will be analyzed statistically and conceptually.

1. Methodology for collection of a good practices in Big Data

The following methodology is based on Guidelines for Collecting, Reporting and Using Data on Innovation, 4th Edition [1].

1.1. Scheme of Big Data good practice

The solution which is based on Big Data can be presented in Fig. 1.

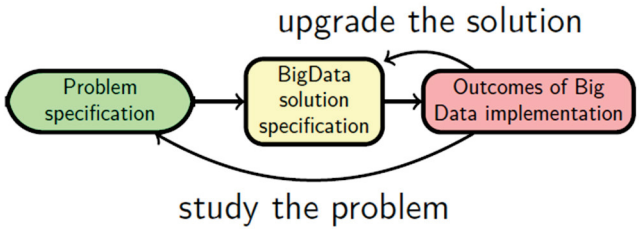


Figure 1. Big Data - good practice

In turn, the Big Data solution specification is presented in the terms of processes:
- Collect / Ingest,
- Store,
- Process/ Analyze,
- Insight / Consume.

1.2. General scheme of Methodology for collection of a good practices in Big Data

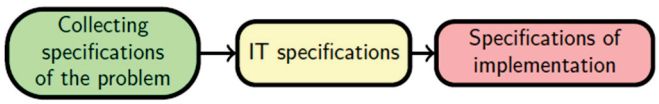


Figure 2. Methodology for collection of a good practices in Big Data

Collecting specifications of the problems causing the development of a good practices in Big Data means the description of the problem, the focus group and the current state of the solutions. We offer the following structure of the specifications:
1. Type of the problem,
2. Who is in the focus? (focus group),
3. How the problem appear?
4. Description of the problem,

5. Current state of the problem,
6. What is a current solution?

IT specifications include the architecture, data, platforms, analytics and machine learning of the Big Data solution. We offer the following structure of IT specifications:

1. Architecture,
2. Data representation,
3. Data processing and quality,
4. Platforms and Tools,
5. Analytics and Machine learning.

Architecture description of the Big Data solution is preferably presented as a process scheme using the following steps: (Ingest, Store, Transform, Analyse, Insight/Application) and designated tools. There a lot of its presentation in a graphical version, e.g., `Ingest Data– >Transform– >Store– >Transform– >Store – >Analyse– >Insight` is presented in Fig. 3.

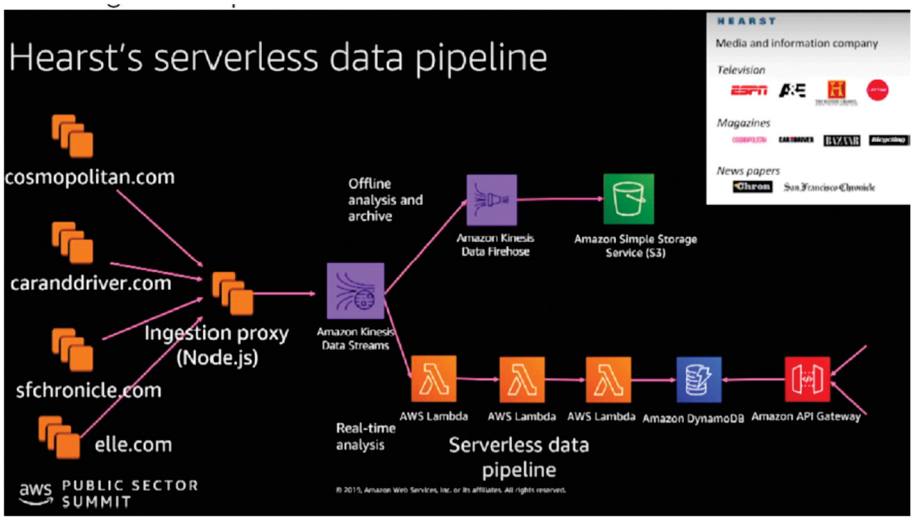


Figure 3. Architecture description of the Big Data solution [2]

Data representation was offered to be investigated as a result of the questionary:

- What is the source of data (repository, service, application)?
- How many data sources are used? Please describe them briefly?
- What is a volume (in GB, TB) of data and its characteristic (streams or blocks of data)?
- What is a data tier? (e.g. records [SQL, noSQL based], files [types], key-value pairs or graphs),
- What tools were used to store data? (Fig. 4 for the preliminary list of the tools).

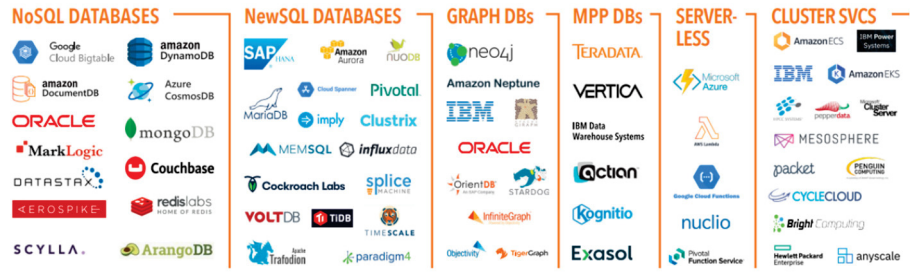


Figure 4. Tools used for storing data [2]

Specifications of the data processing and quality include:

- description of the Velocity of Big Data? (in terms of hot or cold data; it means that its usefulness decrease with time or not); the ways that they should be processed,
- description of the Variety of Big Data (structure of data e.g. records [SQL, noSQL based], files [types], key-value pairs or graphs and initia,l relations (one type of data/ multiple – types),
- description of the Veracity (quality of data - trustworthiness); are the data "clean" and accurate? Do they really have something to offer?
- security feature support in the implementation (what application was used)?
- tools used when process data and to ensure their quality (Fig. 5 for the preliminary list).

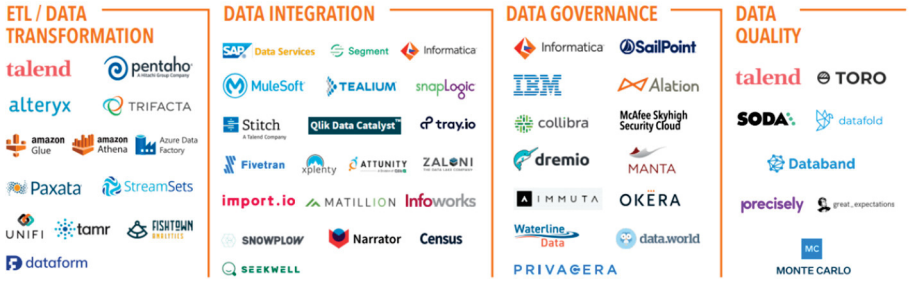


Figure 5. Tools used for processing data [2]

The description of the Platforms and Tools can be covered with the help of the questionary

- What Big Data platform type was used (e.g. server based, cloud solutions, with/without edge computing support or other)?
- What platform solution was used (Cloudera, Aws, Azure, Google platform, other …)?
- What kind of tools/application was used for Data extraction/Ingestion (e.g. Apache kafka, script – node.js)?
- How will the data be accessed? (by put/get , by relations (simple and complex), faceting, search, graph traversal)?
- What type of storage was used (cluster, stream based, data lake, other)?
- What kind of additional applications/tool was used in data storage (e.g. Hadoop, Amazon S3)? (Fig. 6).

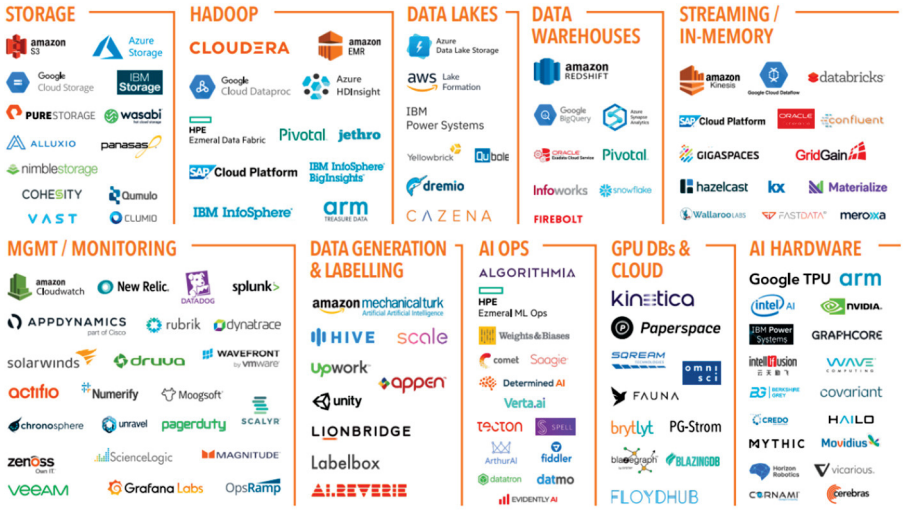
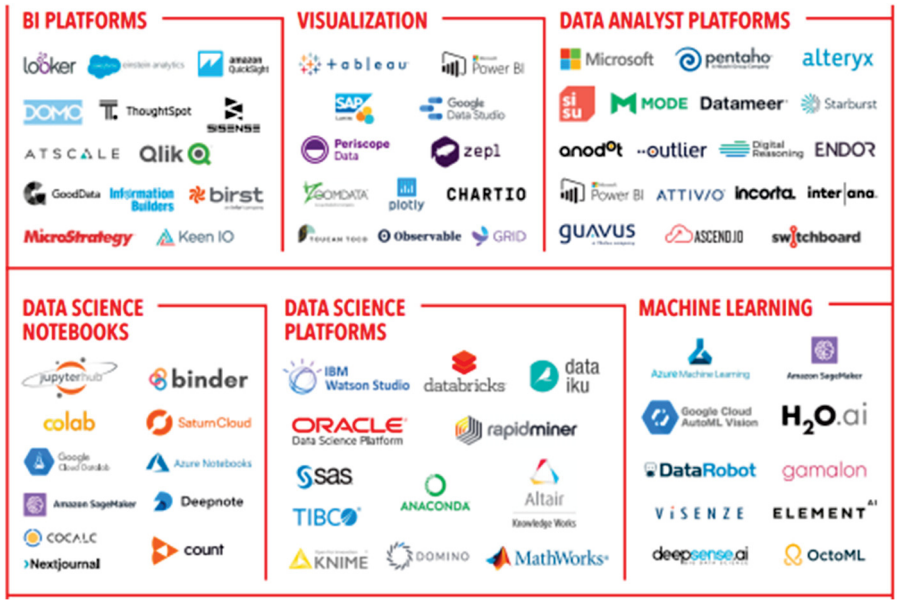


Figure 6. Additional tools used in data storage [2]

Description of Analytics and Machine learning should incorporate:
- kind of analytics performed (statistical, classic machine learning, deep learning),
- kind of applications/tool used in analytics stage (dedicated to a data type / general),
- programming languages used (Python, Java, R, other),
- software used for Analytics and machine learning (Fig. 7 as an example).



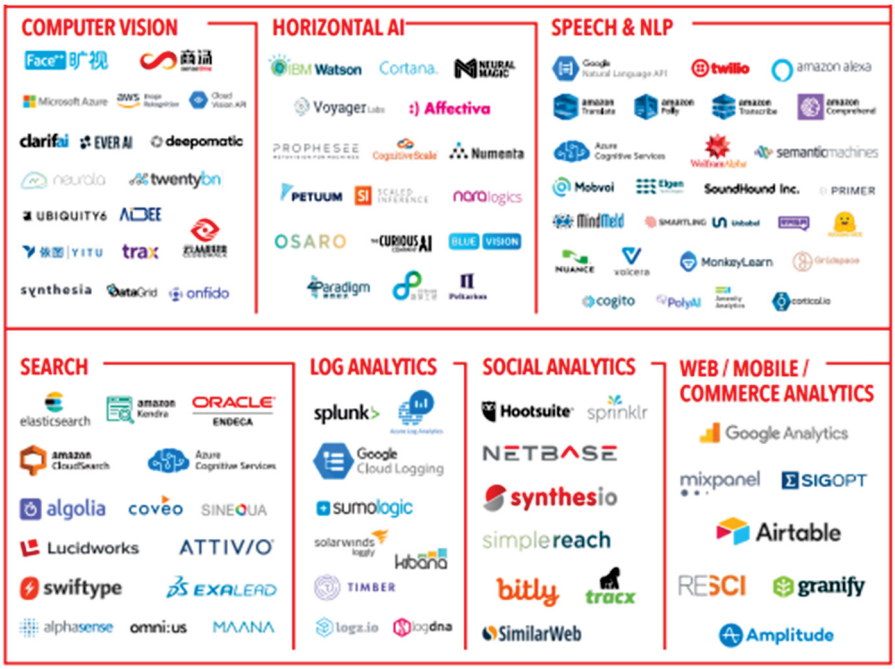
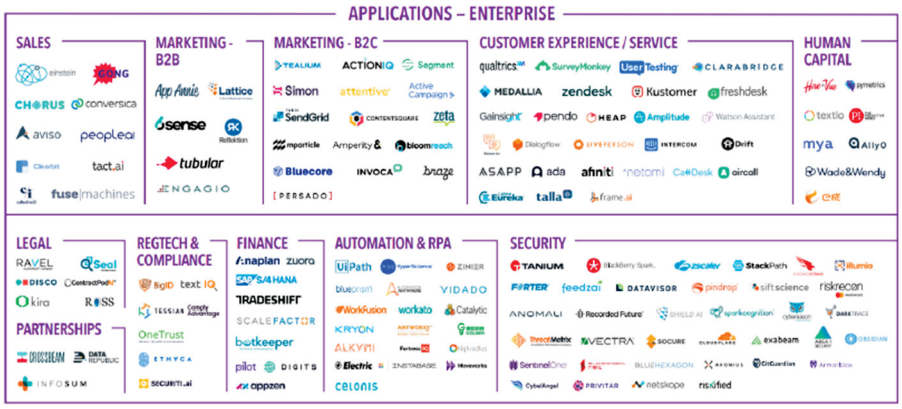


Figure 7. Tools form Analytics and Machine Learning [2]

The specifications of implementation are focusing on the description of the branch of industry, sources or resources and the issues of licensing. It includes the topics:

- the specification of the area of implementation of the Big Data solution (Fig. 8),



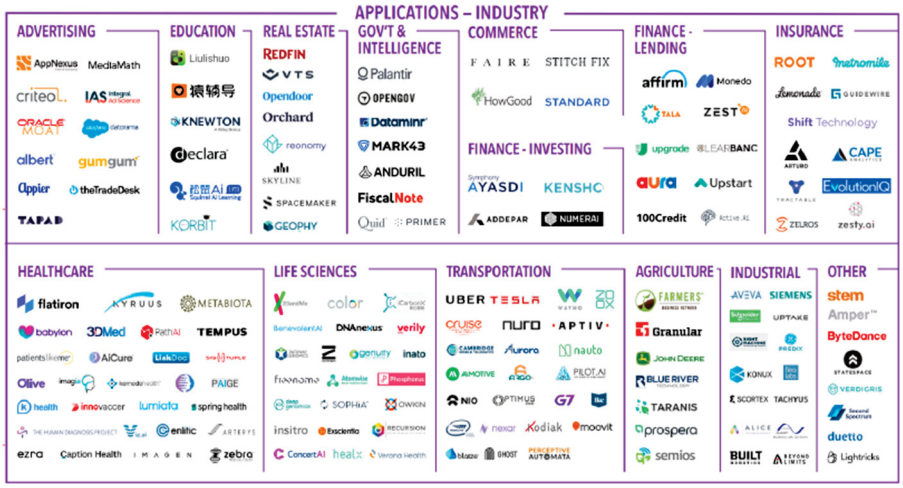


Figure 8. Areas of implementation of Big Data Solutions [2]

- sources and resources of the data used (Fig. 9),

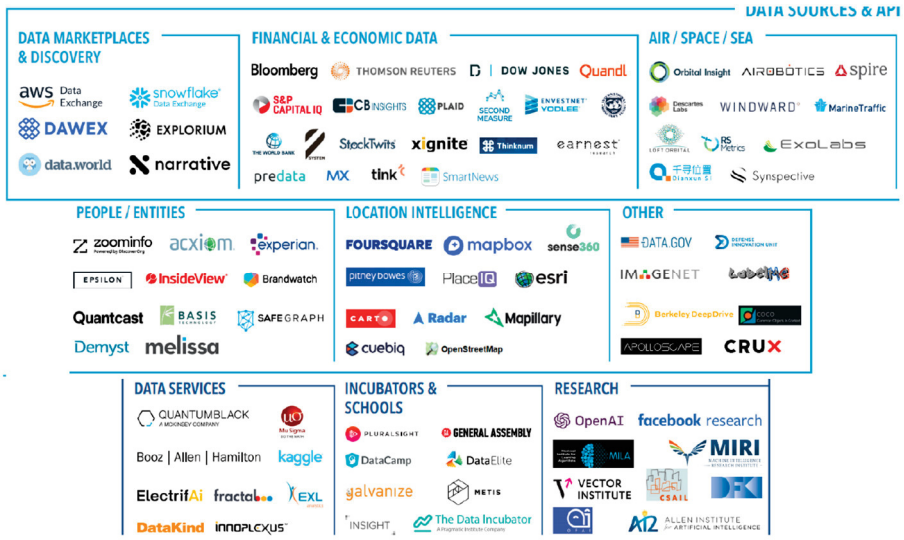


Figure 9. Data source & APIs [2]

- kind of information was retrieved,
- kind of applications/tool was used in data Insight/Consume stage (e.g. Tableau, R Studio, other),
- type of licensing was used for the solution,
- usage of the open-source tools by the solution (Fig. 10).

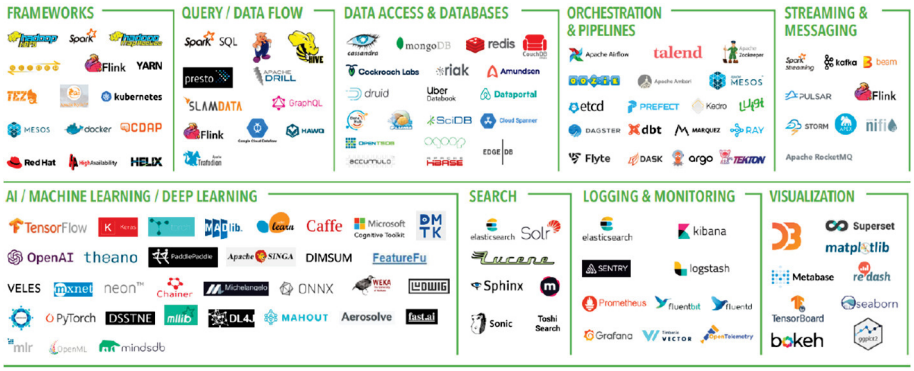


Figure 10. Open source tools for Big Data [2]

2. Methodology of assessment

In the first phase of the project 2020-1-PL01-KA203-082197 "Innovations for Big Data in a Real World" (iBIGworld) under the Erasmus+ program, a study was conducted on the state of the subject area, namely Big Data in a Real World and innovations within it. 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 7 on-line forms with both open and closed questions. The questions consider different questions for job offering, labour market in the field of Big Data, existing training programs and good practices and collecting IT specifications of good practices in Big Data. Also for IT graduates Masters and IT Alumni in Information Systems and Technologies and for Employers to Specifying Graduate Competencies.
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.

Phase 1 was implemented through several on-line forms in different directions, namely:

Research 1: Researching of existing training programs in the field of BIG DATA.
Research 2: Researching of BIG DATA labour market an overview of the demand in the field of BIG DATA.
Research 3: Researching of existing science programs in the field of BIG DATA.
Research 4: Research of the competencies and skills of teachers in the field of BIG DATA (survey for academic lecturers).
Research 5: Questionnaire for IT graduates Masters and IT Alumni in Information Systems and Technologies.

Research 6: Questionnaire for Employers: Specifying Graduate Competencies in Data Science.
Research 7: Collecting IT specifications of good practices in Big Data.

3. The organization of the survey study

The research was conducted by scientists from the 4 countries - participants of the project - Poland, Ukraine, Bulgaria and Serbia.

The surveys were made without the numbering of the questions. This final report contains the conclusions for each report of surveys.

Each survey contains different number of filled questionnaires (Table 1).

Table 1. The quantitative analysis of the surveys filled in

Research	Number of completed surveys
1: Researching of existing training programs in the field of BIG DATA	65
2: Researching of BIG DATA labour market an overview of the demand in the field of BIG DATA	55
3: Researching of existing science programs in the field of BIG DATA	30
4: Research of the competencies and skills of teachers in the field of BIG DATA (survey for academic lecturers)	80
5: Questionnaire for IT graduates Masters and IT Alumni in Information Systems and Technologies	631
6: Questionnaire for Employers: Specifying Graduate Competencies in Data Science	38
7: Collecting IT specifications of good practices in Big Data	15

4. Results

4.1. Results on the research of existing training programs in the field of BIG DATA

The results of the survey have shown information about training courses in the field of Big Data technology and Data Science.
While preparing the courses, the following conclusions should be taken under consideration:

- The survey shows that the training courses in the field of Big Data organized in USA, UK and EU are promoted and accessible for clients for the best.
- We need to focus on the experience in organizing and conducting in training courses in the field of Big Data obtain from USA, UK, EU and other countries (Serbia and Ukraine.

Training courses in the field of Big Data study require primary study in the field of IT. So, the most of training courses are offered within master degree programs.

Short courses are less preferred as the time requirements are not satisfied to present wide range of topics needed for Big Data.
Bachelor and Master programmes in the field of Big Data are preferred in comparison with courses at the academy, since they are advanced continuation of traditional IT courses.
Big Data technology assumes the knowledge and skills of the basic IT technologies. So, the most of propositions of training courses are offered at the advanced level.
Big Data is one of novel technology. Therefore, a lot of persons are novices in this branch, requiring the courses at the level for beginners.
The are definitely less proposition of training courses at the level of intermediate and managers, which shows the great opportunity and perspective when developing such level courses.
When developing the name of the planned course in the field of Big Data, one should bear that the training course should include the term "Big Data" clearly.
It is desirable to show in the name of the course the connection with data science and analytics.
Some in formation on the level of the program (e.g. bachellor, master or at least advanced) should be indicated in the name of the training course.

The prerequisites of planned course in Big Data have to include:
- the training course in Big Data has to assume certain level of programming skills; requirements for enrolling should assume college level of knowledge of Python, R, SDS, etc.;
- requirements for enrolling should assume college level of calculus and algebra;
- skills of using databases (SQL and NoSQL) are to be required for enrolling for the Big Data course;
- the ability of algorithmic approach (e.g. data mining) to solve the problems dealing with the data is to be required for enrollment for Big Data course.

The planned course in the field of Big Data has to include the following topics:
- system architecture of large-scale system using Big Data;
- Big Data platforms: Apache Hadoop Ecosystem and Their Components;
- distributed streaming data-flow computing with Big Data (Apache Flink, Apache Storm);
- cloud computing (Hadoop Mapreduce and Apache Spark);
- software bus and stream processing for Big Data solutions (Apache Kafka);
- analytical processing of Big Data (Apache Spark);
- dataflow between systems; using (Apache NiFi);
- business analytics based on Big Data;
- architecture of microservices for Big Data solutions;
- algorithms of analytics and machine learning for Big Data;
- intelligent (e.g. Bayesian) analysis of large-scale data;
- deep learning for Big Data;
- forecasting and risk assessment for large-scale data;
- high performance analysis of large-scale data (e.g. world econometric, geospatial);
- visualizing Big Data.

The advertising of the training courses offered in the branch of Big Data should include some extended descriptions of the topics covered in the course.
When advertising training course, the extended descriptions of the topics covered should include both direct Big Data topics and specific topics related to the ways of processing large volume of data, namely, taken into account data structure, large scale, efficiency, parallelism, etc.

The topics of visualization and presentation of Big Data are of important.
The training courses should include good practices of processing Big Data gathered form real application.
The topics of using analytics tools should be included in the training courses in Big Data.
The developing and using data storages based on e.g., Data Lakes, Hadoop and others have to include in the training courses.
The data security was mentioned as the less important when considering the related topics of Big Data.
The highest interest was noticed to receive a full academic education. A master's degree was the most preferred, but a bachelor's degree was also in high demand.
It was noted that certificate and postgraduate diploma are the less interest for professionals.

Certificates (training and masters) and credits courses were the least popular among existing training courses in Big Data:

- The most preferred form of knowledge assessment are: exams, projects, assignments and obtaining the appropriate numbers of ECTS points.
  MS thesis, course work are ranked in the middle.
  The less popular form of knowledge assessment are: practical exams, seminars, laboratory, homework, presentation, online forms, podcast, quizzes, programming exercise session, self-study and team projects.
  When planning and developing training courses in Big Data, one should bear in the mind the following aspects of the duration of course:
  The most of courses are offered for one semester, but also popular are courses for one year and four years.
  Trainings for 1,5 year, 1 day,13 lectrures, 14 classes, 14 weeks, 2years, 220 hours, 3 days, 3 months, 3 semesters, 4 semesters, 4 years, 6/8 weeks, 8 months, 8 weeks, 9 months, 4 months, 12 months, 1,5-3 years are not so popular.
- The highest price is 35200 Euro, but there still exist many Big Data courses organized by free.
  On average, the course fee is 6984,182 E

4.2. Results on the research of BIG DATA labour market: an overview of the demand in the field of BIG DATA

The results of the survey shows that the profile of a Big data expert searched on the market is connected mostly with IT specialists with additional features. While preparing its profile and a course the following conclusions should be taken under consideration:
There is a need for Big Data employees on the market.

The work place is in EU countries and the UK.
The Big Data knowledge is a supplementary one to Data Scientist, Data Engineer and Data Analyst occupations.
Big Data is part of IT science and data analytics.
The difficulty of the subject requires full engagement.
The general (not dedicated for specific area) skills are in demand.
The Big Data course should be naturally offered as a part/extension of Computer science study.
Can be offered as skills training as part of a post diploma course.
Both inexperienced and experienced Big Data specialists are needed.
Course should be focused on the basics.
The course should cover some practical experience (e.g., project).
Python language should be incorporated as required skills.
The R language, Java and SQL should be treated as part of the training program or perquisites for the course.
Soft skills covering Communication, Critical thinking and collaboration must be developed within a course.
The skills: working with tools, reskilling and Creativity should be supported in the course.
The prepared curriculum should focus on processing Big Volumes of data and knowledge of appropriate algorithms to tackle them.
The course should provide description of various analytical techniques, processing SQL, NoSQL datasets, visualise data, using analytical platforms and process large datasets.
The open source tools should be proposed that are available on cloud platforms.
The full pipeline should be presented.
The stress should be put on data analytics solutions.
The streaming and cloud technology should be added if possible.
The basic knowledge of pipeline elements is essential.
The Hadoop ecosystem, AWS and Azure are preferred technologies.
Big Data is developing into an analytics area and processing area, while the technical elements are automated.
The job market is looking for issue solvers rather than strict technicians.

4.3. Results on the research of the existing scientific programs in the field of BIG DATA

The survey results show that:

The difficulty of obtaining detailed information about scientific and commercial projects in the field of Big Data suggests that the field of Big Data is in a state of development. Projects are trade secrets.
The analysis of the survey results showed that the researches covered different areas of Big Data applications. There are projects supported by the European Union within the framework of the program Horizon 2020.
It is necessary to encourage universities to do research. Universities should become centres of technology transfer to industry.

Many business representatives note that difficulties in implementing Big Data projects are associated with a shortage of specialists - data scientist, data engineer, data analyst. The rate of return on investments in Big Data directly depends on the quality of work of employees engaged in deep and predictive analytics. The variety of goals and therefore the content of Big Data projects indicates the need to develop educational programs and train specialists who are able to solve problems in the field of Big Data. The variety of data sources requires the development of ways to collect and store them. These issues should be focused on the developers of educational programs in the field of Big Data.

Meeting the business needs for Big Data solutions requires expertise, technology, tools, infrastructure, and new challenges. It is obvious that the training of specialists in the field of Big Data is an urgent task.

To carry out scientific, educational and commercial projects, universities must have the appropriate resources: personnel, equipment, finances and time. Universities will be able to achieve the desired results in the structure of educational ecosystems, research and production clusters, organized according to the quadruple helix model. The development of knowledge-intensive industries is necessary for a high-quality innovative breakthrough in the creation of a competitive high-tech economy. It is universities that can become points of growth that ensure the innovative development of the economies of countries. Indeed, the specificity of the activities of universities makes it possible to effectively integrate personnel training, research, and commercialization of the results of intellectual activity. Analysis of Big Data projects has proven the ability of universities to generate knowledge in the form of disruptive scientific research.

To achieve the set goals and obtain the results declared in the project, resources are needed: people, time, finances. An educational project aimed at developing educational courses on Big Data should contribute to understanding what content should be included in programs and what competencies should be formed in students. The competencies of students with a Big Data education should include the ability to use tools to handle Big Data sets. In this case, Hadoop can be one of the first tools that students will master.

Teamwork, the ability to communicate, knowledge of conflict management are the skills that students need to develop, including the relevant disciplines in educational programs based on Big Data

It was not possible to obtain reliable information about the structures, composition, data types of the project datasets analyzed as a result of the review of project sites. As a rule, data is a commercial secret of projects.

The participation of partners in projects contributes to the success of projects. Projects funded by EU programs (Horizon 2020) require partners to participate in projects. As a rule, the staff of the companies carry out commercial projects. Sub-contractors may be employed to carry out highly specialized project tasks. Commercial and grant-financed projects can be considered more successful, probably due using of more flexible personnel management schemes. The increase in the number of project participants should be justified and determined by the complexity of the task being performed, as well as the duration of the project. The project partners partially act as the guarantor of the expected success.

An academic degree is not required for a project leader. But its presence is desirable, since it acts as a guarantee of high-quality project implementation without violating

deadlines. About half of the project managers are professors or have a Doctor Sciences degree. Commercial and grant-financed projects can be considered more successful, probably due using of more flexible personnel management schemes. Project managers with sciences degrees have certain advantages since they have fundamental and technical knowledge, team building and management skills already. High professionalism and organizational skills are needed a manager to lead to the success of the project.

4.4. Results on the research of the competencies and skills of teachers in the field of BIG DATA (survey for academic lecturers)

The survey results show that:

The intellectual potential of teachers in the field of Big Data needs to be developed. The main problems in the development of the Big Data direction are the lack of qualified personnel, the lack of sufficient implementation experience, as well as the high cost of solutions.
Teachers of partner countries' universities are creative people who are able to improve their qualifications and gain new knowledge, including in the field of Big Data. It is necessary to encourage and motivate teachers to improve their competence in the field of Big Data.
It is necessary to expand the teaching of Big Data courses in universities and provide teachers with the opportunity to gain experience and improve their knowledge and skills.
Most teachers do not have pedagogical experience, and the corresponding competencies have not been formed when conducting courses on Big Data. Most teachers understand that teaching student relevant Big Data courses in accordance with the requirements of the IT market is possible only with the involvement of external experts from the IT industry.
It is necessary to expand the possibilities of cooperation universities with enterprises in order not only to attract specialists to teach disciplines but also to involve teachers in project work. In addition, universities should create opportunities for teachers to take thematic courses at the request of the teacher.
It is necessary to improve educational programs in universities constantly, taking into account the requirements of the IT industry and modern trends in IT development.
Teachers would like to have an online platform highlighting the results of work in the field of Big Data, carried out by partner companies of universities.
The best way to improve your skills is to participate in open source projects. By participating in various projects, developers can improve their skills and gain inspiration and support from like-minded people.
Participation in Big Data Part 3 (employer-sponsored) courses to expand teachers' knowledge is very important for teachers.
For systematic work in the field of Big Data, teachers need to create opportunities for participation in conferences, for the systematic solution of applied problems in this area in the framework of scientific and real projects.
Teachers would like to use Big Data analytics and machine learning algorithms to solve real-world everyday problems. The popularity of these areas can provide breakthroughs in startups, victories in student Olympiads etc.

We need a program to harmonize relations between IT companies and universities to attract teachers to participate in real commercial projects. It is necessary to popularize the direction of research related to Big Data, to motivate teachers to master Big Data technologies, to create conditions for the participation of teachers in projects and research related to Big Data, to ensure the availability of technical teaching materials on Big Data.
A strategy is needed to develop teacher competencies in Big Data, create opportunities, fund research, and test its results in the real manufacturing sector.
To increase teacher participation in Big Data research, it is necessary to create conditions to motivate and incentivize IT companies to attract teachers to work on projects. For this, it is necessary to legalize the part-time work of teachers in companies.
Teachers lack faith in their ability in terms of their willingness to teach Big Data. It is necessary to motivate and stimulate the desire of teachers to teach students Big Data.
A modern university should be an educational ecosystem in which the university, companies interact, in particular the IT industry, accelerators, business incubators, various funds, etc. technologies in the educational process.
Practice-oriented learning takes into account the demands of the labor market, where Big Data solutions dominate all industries. The professions Data Scientist, Data Engineer and Data Analyst are among the scarcest. The introduction of Big Data courses in the educational process of universities should become one of the priority areas of activity.
The solution of the tasks set before the universities requires the pooling of resources and potential of the academic community (universities, faculties, departments, laboratories), business organizations, government bodies and society. The introduction of models of the triple and quadruple helix, the creation of knowledge clusters will solve the problems identified by the respondents.
To make learning more focused, students should participate in the implementation of research projects with both universities and companies. For this, it is necessary to use the unique equipment of the companies. Given the growing demands of potential employers, it is necessary to involve employers in the development and improvement of training programs.

4.5. Results from Questionnaire for IT graduates Masters and IT Alumni in Information Systems and Technologies

The interest of IT graduates Masters and IT Alumni in Information Systems and Technologies for the online survey, in the scope of the iBigWorld project, was impressive, with 621 participants, mostly from the partners' countries. That interest demonstrates the clear need of the target population for Big Data contents as well as Big Data trainings and courses. Respondents were mostly young people (70% younger than 24 years) still at universities or working for a short period of time.
Survey respondents were mostly student population engaged in Bachelor and Master programs of different kinds of IT studies, proving that the objectives of the project and target groups were set correctly. Good foundation of the iBigWorld project and the growing need for courses dealing with Big Data are also confirmed by the fact that the majority of students interested in the survey on the Big Data, do not posses the knowledge about any relevant course at their university. Surveyed students

demonstrated large interest in Big Data content and new courses whereby their opinion is that the developing course should lean heavily on practical implementation with laboratory work, students' projects and internships. Having almost all of the IT students interested or neutral regarding studying Big Data course, is in line with assumption of the project that there is a gap in southern Europe regarding the Big Data and other digital skills and a need for new initiatives to ensure that these skills are adequately promoted in the curriculum, in teacher development, in assessment practices and in learning content.

Survey showed that partner countries of iBigWorld project should focus on implementing Big Data content into study programs by either introducing completely new courses or by innovating existing courses. Improving material component (equipment, laboratories) is secondary, but also important factor. Presenting Big Data research in a form of website should be one of the most important goals for partner universities, because that kind of informing is the easiest way to motivate and include students and other target groups in Big Data topics. Large interest of the students in Big Data issues, despite their low level of insight adds to the project goal to identify the underrepresented skills, the rationale behind the phenomenon of talented people who lack the traditional credentials to land a good job and the Big Data with the most pressing needs. Answers obtained by the students interested in taking part in Big Data Courses showed that designed Big Data courses and trainings should be attractive if they are heavily based on techniques of collecting data from different sources as well as on analytics of those data, i.e., machine learning and data mining techniques.

Almost the same number of questionees were currently working or were without a paid job. This proportion allows needed diversity for the survey analysis and enabled analysis of the needs of working people also, not only the students. Answers provided by working population showed that designed Big Data trainings should be tailored particularly toward people working in the private sector (by content, time of lectures...). These people are mostly motivated to improve their competencies and to learn more about Big Data. Proposed Big Data courses should target younger people with less working experience because they are more motivated and still interested in learning new topics, for which they feel that could help them improve their careers.

Specific attention in the survey was paid to the workers in the IT industry. Results showed that designed Big Data trainings should be heavily oriented toward Software developers because they are making the majority of the target group for developing digital competencies and skills in the area of Big Data. During designing Big Data trainings, project partners must have in mind that workers in IT sector either work at positions where there are no requirements regarding their education or only bachelor diploma is needed (together 80%). These data are consistent with the lack of IT experts in the market. So, the most logical choice to place Big data courses would be at Bachelor studies or as part some specialization programmes. Large majority of workers in IT sector either have no experience or very limited experience (up to one year) in Big Data field meaning that designed courses have to be at the basic level providing all the necessary knowledge for dealing with Big Data.

Large majority of working respondents (87%) answered that Big Data issues are important or somewhat important to their job and that Big Data contents would be valuable for their future career development, proving the foundation of iBigWorld project and importance of introducing structured trainings in Big data topics. Developing of digital and soft skill becomes more and more important in today's job

market. Young people recognize their importance as well as their employers. Any designed training should also focus on developing soft skills, especially those related to teamwork, communication and time management. Most valued professional competences which a specialist in Big Data should have and which should be taught during Big Data trainings is Effectively use variety of data analytics techniques (Machine Learning, Data Mining, Prescriptive and Predictive Analytics). Analysis of other survey questions also demonstrate that Machine Learning and Data Mining are the most popular aspects of Big Data and data analytics.
Evaluating competency training demonstrated that IT workers highly value both soft competencies like Development of logical, algorithmic, systems thinking and Ability to adapt to changes in the IT market, as well as professional skills like Knowledge and skills in software development and Acquiring skills in managing IT projects. Difficulties which IT workers encountered during their career once again confirm the proclaimed goals of iBigWorld project. Namely, IT workers answered in large majority that they encountered situations where Knowledge was required, which they did not posses or Competence was needed, which they did not have. Both problems should be tackled by designing adequate Big Data trainings in cooperation with business sector. Big Data trainings and courses and guidelines for their design have to take into account desirable professions for which the respondents are working or want to work (in accordance with the list of the European framework of IT competences) and to be tailored specifically around them, especially professions like Software Development, Project Manager, Technical Specialist, Systems Administrator, and Systems Architect.

4.6. Results from Questionnaire for Employers: Specifying Graduate Competencies in Data Science

The results of the survey shows that the employers are interested in BD and DS graduates and appreciate their skills and expertise in their business activities related to Big Data. For achieving the project goals and preparing the appropriate course(s) related to BD and DS that would satisfy employers' needs and expectations the following conclusions are important based on this questionnaire:

They clearly express their interest to improve their business toward Big Data – related domains and activities.
Most of the organizations offer or plan to offer jobs related to BD and DS and prefer to employ graduates with at least one year of experience, up to three years and more.
They require some BD-related skill set from their potential employees, various abilities in design, implementation and deploying BD solutions, as well as their maintenance.
They are mostly not satisfied with the knowledge and skills of master students and graduates but will appreciate the improvements of the BD-related curriculum and courses, as well as development of appropriate courses within organizations themselves.
The organizations also require the competencies regarding skills and expertise in various areas, such as: Business Continuity and Information Assurance, Systems Development and Deployment, Data, Information and Content Management, Enterprise Architecture and IS Management and Operations.

The part of the questionnaire also emphasized the need for general IT and social/personal competences not particularly tied to BD and DS.
All mentioned and required competences should be properly analyzed and considered when preparing Big Data course(s) within the project, aside from technical BD and DS skills and competencies.

4.7. Results on the collecting IT specifications of good practices in Big Data

In this research are looking at 15 cases/solutions that use Big Data. The most of described cases/solution in the field of Big Data come from 10 non-EU countries, and only 5 are EU members. It can be seen that the majority of organizations use sources that are partially open-source, followed by the company using open-source solutions. A little more than half part of the projects described used research companies specialized in the field of artificial intelligence for open-source resources. Fewer companies use Data services (like Quantum black, Kaggle, ElectrifAI).

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 a different kind of information.
The most used tool for working in these projects is Jupiter, followed by Python and R studio.
All 15 projects include almost all of the listed typical important steps, for solutions, using Big Data - ingest, store, transform, analyse and insight/application and designing tools.

The projects use several sources for data accumulation. Most of the data is accumulated from databases, from services and sensors, and a little from the use of applications, weather data or semi-structured files.
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.
The largest share is represented by data in the form of records in noSQL databases, followed by records in SQL databases. In the cross-analysis between the different data types and their volume, it is noteworthy that most of them are records in SQL and noSQL databases, and their size is over 1GB.
Regarding the type of platform used for Big Data - the considered projects are implemented on 2 types of platforms - dedicated server and cloud solutions. However, in a cross-analysis, it can be seen that some projects use both types of Big Data platform.
Various applications have been used to extract data from these projects. In some projects only one application was used, in others several were used.
The most common data retrieval tool is the Apache Kafka software platform - it is preferred in most 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.
The study also traced what type of data storage was used in the described projects.

Many 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.
Of the additional tools and applications for data storage and management, the most usable is HADOOP. The platform 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.
The preferred tool for information processing are Data Science Platforms - in the 40% of the projects. It is followed by Machine learning through Azure Machine learning, DataRobot and others, preferred in 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.
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 20% of the projects - Apache Spark, Python scikit-learn. In 13.3% of the projects selected only Apache Spark of machine learning and other 13.3% is selected only R Studio.
About the date sets used - goals 66.6% use free date sets, while others 33.3% - used paid.

5. Conclusions

So, the work offered the methodology for collecting good practices in the field of Big Data. Moreover, this technique has been applied to designing the research and analyzing the results. Particular attention was paid to the hard and soft skills as well as the topics that should be covered by the Big Data training course.

REFERENCES

1. Website: *https://www.oecd-ilibrary.org/oslo-manual-2018_5j8p8jmbxgwb.pdf?itemId=\%2Fcontent\%2Fpublication\%2F9789264304604-en&mimeType=pdf.*
2. Website: *http://mattturck.com/wp-content/uploads/2020/09/2020-Data-and-AI-Landscape-Matt-Turck-at-FirstMark-v1.pdf.*
3. Top 10 Big Data Tools that you should know about – DataFlair.
4. What is Big Data - Importance and Use Cases - DataFlair (data-flair.training).
5. Public sector sumit 2019 AWS: *https://www.youtube.com/watch?v=MotN5f6_xl8.*
6. Google cloud process 2019: *https://www.youtube.com/watch?v=GRP-cGbJSCs.*
7. Website: *https://www.youtube.com/watch?v=LZgqGtOLq80\&t=2300s.*
8. Website: *https://www.youtube.com/watch?v=DMck8_5rPhU.*

9. Best practices:
https://www.youtube.com/watch?v=hsq4s_l9ZDM\&feature=youtu.be.
10. Key trends article: *https://mattturck.com/data2020/*.