DOI: https://doi.org/10.53052/9788367652018.12

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

PRZEWODNIK DLA NAUCZYCIELI A4.1 W RAMACH PROJEKTU ERASMUS+ IBIGWORLD: PODEJŚCIE DO TWORZENIA KURSÓW SZKOLENIOWYCH OPARTE O PRAWDZIWE PRZYPADKI

Streszczenie: W wyniku zorganizowanego szkolenia w ramach projektu Erasmus+ nr. 2020-1-PL01-KA203-082197 pt. „Innowacje dla Big Data w realnym świecie", stworzono wytyczne dla nauczycieli. Ogólnym celem tego kursu jest poprawa zdolności studentów do rozpoznawania koncepcji biznesowych i trudności związanych z przepływem pracy Big Data, a także studiowanie i znajdowanie kreatywnych rozwiązań problemów na dużą skalę. Trenerzy powinni poprowadzić kursantów do procesów biznesowych związanych z Big Data, aby studenci mogli je rozpoznać i wiedzieć, jak mogą być przetwarzane.
W tym celu opracowano wytyczne, które pomogą nauczycielom zorganizować kurs oparty na rzeczywistych przypadkach, aby przygotować takich specjalistów.

Słowa kluczowe: przewodnik dla nauczycieli, szkolenia, Big Data, iBIGworld

TEACHER GUIDE A4.1 FOR IBIGWORLD ERASMUS+ PROJECT: APPROACH TO CREATING A BIG DATA TRAINING COURSE THOUGHT REAL CASES

Summary: As a result of an organized training course within the Erasmus+ project no. 2020-1-PL01-KA203-082197 entitled "Innovations for Big Data in a Real World", a Teacher Guideline was created. The overall goal of this course is to improve students' capacity to recognize business concepts and difficulties associated with Big Data workflow, as well as to study and find creative solutions to large-scale problems. The trainers should guide the trainees

[1] University of Library Studies and Information Technologies (ULSIT), Sofia, Bulgaria: (i.kostadinova, g.dimitrov, v.totev, e.kovatcheva)@unibit.bg
[2] Department of Computer Science and Automatics, University of Bielsko-Biala, Poland: (vmartsenyuk, tgan)@ath.bielsko.pl, damiangrygiercz@gmail.com, palka99kacper@gmail.com, mateuszdamek@onet.pl, wk054421@student.ath.edu.pl
[3] University of Niš (UNi), Nis, Serbia: dejan.rancic@elfak.ni.ac.rs
[4] Taras Shevchenko National University of Kyiv (TSNUK), Kiev, Ukraine: oleksiibychkov@knu.ua

to business processes related to Big Data so that the students can recognize them and know how they can be processed.
For this purpose, a guideline has been developed to guide teachers on how to organize a course based on real cases to prepare such specialists.

Keywords: teacher guide, training, Big Data, iBIGworld

1. Introduction

Big Data and its analysis techniques are at the center of modern science and business. Extracting valuable knowledge from massive quantities of data is complicated due to the sheer volume of data generated every day. It becomes difficult to capture, form, store, manage, share, analyze, and visualize meaningful insights from the data. The domain engaged in solving these challenges is collectively described as "Big Data", but how to deduce valuable information out of large sets is an area of particular interest in "Big Data Analytics" (BDA) (Fig. 1).

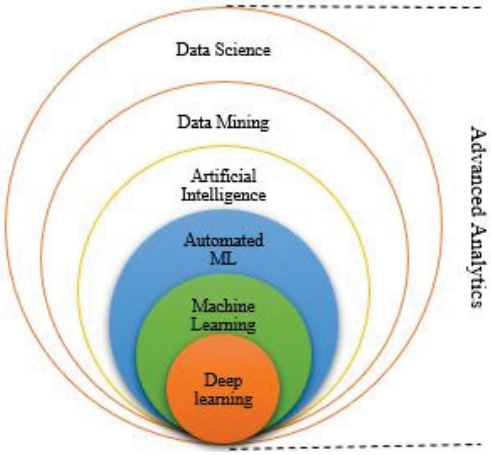

Figure 1. Knowledge Discovery Paradigm

Big Data processing is a complex activity involving different experts or an expert with various competencies, and the art of value extraction is the heart of Big Data Analytics. The last is a whole scientific palette of advanced methods applied, all or selected, depending on the defined problem from different fields to unlock valuable insights.
Analytics is a team sport. Data needs to be located and cleansed, models have to be created, tested, monitored, and updated. All this requires teamwork.
Indeed, although analysis and analytics are closely related, there are some subtle differences between the terms that impact the design of Big Data solutions.
Working on Big Data projects is impossible without team collaboration, especially in the Analytics stage, as it interlinks all processes, depends on them, creates insights and responds for decision-making. Each activity is essential.
More or less, all of the discussed methodologies can be adapted for the purposes of different data-driven projects. Which of them to use depending on the particular

analysis goals, knowledge of experts, V's characteristics of data available, and accessible technologies?
As a result of an organized training course in a iBigWorld: Innovations for Big Data in a Real World (Erasmus+ project 2020-1-PL01-KA203-082197) [1] - competence-based learning in Big Data, a Teacher Guideline was created. The overall goal of this program is to improve students' capacity to recognize business concepts and difficulties associated with Big Data workflow, as well as to study and find creative solutions to large-scale problems. The trainers should guide the trainees to business processes related to Big Data so that the students can recognize them and know how they can be processed.
Lately a lot of attention was paid to the good practice in the development of the teacher's guide [2-4]. It was successfully applied when creating teacher guide for the computer science and data science courses [5-9].
The given teacher guide is structured with the structure due to the following Sections.

2. The Course's Goals through a prism the teachers

The overall goal of this program is to improve students' capacity to recognize business concepts and difficulties associated with Big Data workflow, as well as to study and find creative solutions to large-scale problems.
Participants will be able to handle Big Data innovation management projects that develop answers to existing challenges after completing the course.

➢ Knowledge

Participants will be able to: Competently explain basic innovation management principles;
Demonstrate grasp of if issues of working within a live corporate setting; Competently explain how to use various innovation management tools to produce solutions at the conclusion of this program.

➢ Skills

Participants will be able to: Develop a project plan to tackle a simple business problem; Research and produce a coherent presentation around a proposed solution, and Work with a broad group of individuals with various skills by the end of this program.

3. Determining the level of students and selecting suitable students

The course is designed for final-year undergraduates, first-year master's students, and second-year master's students. This program is especially suited to students who have no prior business experience or education.

4. Teachers training course details - training structure

The 120-hour training course is broken down into 3 stage contain 12 Units (Fig. 2).

Stage 1: Big Data Fundamentals
1. Introduction to Big Data Science and Systems.
2. Big Data Pipeline.
3. Data Mining.
4. Technologies for collecting, storing and managing Big Data.

Stage 2: Big Data Ecosystem and Tools
5. Hadoop Ecosystem.
6. Databases for Big Data.
7. Big Data Analytics Tools.

Stage 3: Big Data real-world use cases
8. Data Acquisition. Workspace. Data Analysis.
9. Data Tidying and Cleaning, Data Visualization.
10. Exploratory Data, Data Analysis.
11. Classification methods, Neural Networks.
12. Statistical Methods.

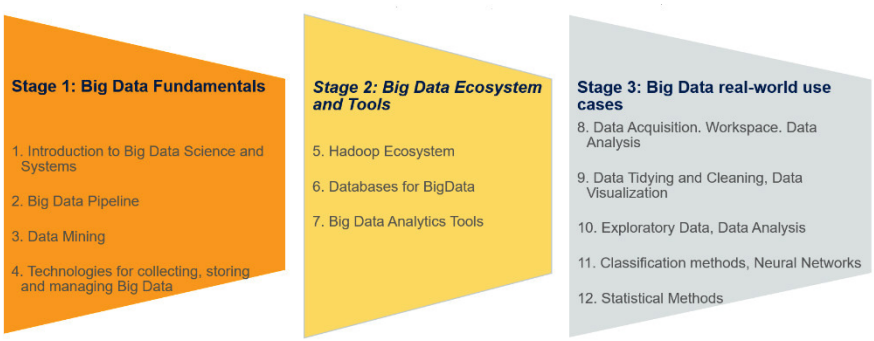

Figure 2. Training course structure

5. Defining the training methods. Training course details (activities)

In the the course are using several learning methods, the basis of which is self-learning. Each of the unit contains following activities (Fig. 3):

- Lectures (presentations and lecture notes),
- Demonstration of real world use cases (presentations),
- Practical tasks in a team (practical exercises),
- Q&A Sessions (questions and discussion),
- Learning Scenarios (branching scenarios).

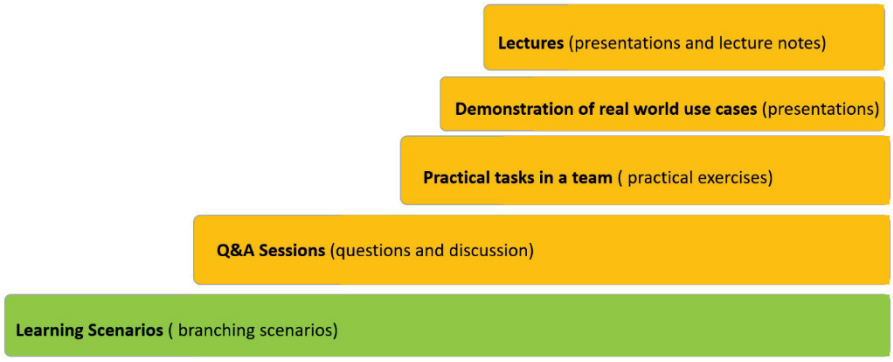

Figure 3. Training methods in the course

Rather than academic theory, the course content is practical and based in the real world Big Data use cases. The course is developed with student participation, expert facilitation, and the application of new approaches and techniques in mind. The training program is designed to match the skillsets of the most successful practitioners.

6. Defining the training techniques

The following strategies should be included in the instructional methods. The trainers are not limited to these tools and approaches, and they are encouraged to try out they own concepts:

- Energisers, icebreakers, and games,
- Seminars,
- Self-directed education,
- Self-study and group conversations,
- Individual and group contemplation, as well as experience sharing,
- Research studies,
- Role-playing exercises.

The following procedures should be included in the participants' learning strategies:

➢ Doing-based learning:
- inquiry-based learning and,
- problem-based learning,
- Competency-based education.

➢ Self-directed education

➢ Action and experience learning

7. Programme Structure

The general structure is centred on students and designed to meet the demands of diverse businesses. Students should given a framework of theories, ideas, thoughts, and concepts to build understanding through a series of workshops. The students then

should work with firms to establish theoretically sound new methods of doing things within that business. As they attempt to put the theories into reality, students must be asked to examine their veracity and value.
The program, which was created using quality assurance descriptors, focuses on the participant's overall growth:

Understanding and knowledge:
✓ of theoretical perspectives, methods, and techniques of innovation management; key elements of success when establishing an innovation strategy; critical components of the innovation management process; financial and risk evaluations of an innovation strategy.

Build a cognitive abilities:
✓ ability to analyze and synthesize information from multiple sources to reach justifiable conclusions; Use conceptual skills to create and implement decisions; Research and develop an innovative solution to a problem; Evaluate the relevant skills needed to manage innovation at various levels; Identify and evaluate elements of an innovation strategy;

Build a soft talents:
✓ enabling students to demonstrate their ability to manage small-scale projects; create a basic project management chart; lead a small group; collaborate productively; communicate effectively with others

8. Tasks to be solved by the teacher

Identifying a requirement:
Make sure there is enough interest in the program for training trougth with real cases. Trainers, participants (students), mentors, and business owners (managers) are the four groups that must be involved in this program, each playing a particular function. One of the first things you should do is make sure you have a trainer on hand.
You should then determine whether can you recruit enough participants for the program in order to trial it first, and then follow your organization's internal procedures to add it to the curriculum as a full-time course.
Recommend is that is put out a call for potential participants to indicate their interest. A minimum of 12 people is recommended. This will allow to work in at least three groups of four people to come up with a solution to the business challenge.
The number of the groups can vary, but they should preferably be between 4 and 5.
It's also a good idea to gauge mentors' and business owners' (managers') enthusiasm, as they play an important part in the program. That can do this by putting out an expression of interest request or approaching local company owners to see if they are interested in participating.

Organizing the meet and training:
After the determined a need, it will be want to make plans for the occasion. If the teacher knows the final number of participants, he will need to find a location to hold

the face-to-face portion of the event. Depending on the participants' knowledge and skill levels, he'll need to run 7 or 8 face-to-face sessions lasting 4 to 5 hours each. It is recommend doing a face-to-face session once a week to allow participants to conduct the necessary research and self-directed learning to satisfy the program's learning objectives.
It is encourage that the teacher is evaluate the material and then add to it, even though there is enough to run the application. Examining the materials and developing a lesson plan to identify aspects that require more attention is always a good idea.

9. What the learning process should be?

This teacher guide aims to develop a case-based training course that is interactive, practical, participatory, and emerging. (Fig. 4).

Figure 4. What the training process should be

10. Structure of themes of Big Data Course

The training course is organized in 3 stages with several themes in every stage (Fig. 2). In Fig. 5 below is shows the structure that is assumed for each of the topics covered in the Big Data training.

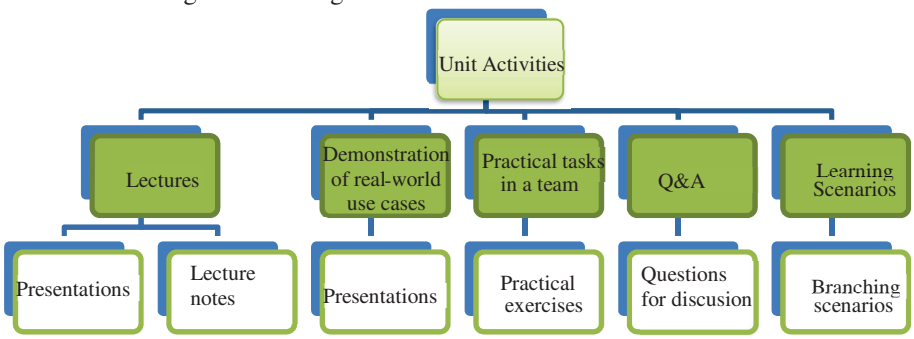

Figure 5. Unit Activities

10.1. Topics on Big Data Fundamentals

The first part of the course is focused on introducing to Big Data Science and Systems. It covers the basic topics that will re required in the next units. The objective is to introduce the main terms and to show their usage. On the other hand, we characterize the basic information processes related to Big Data. Namely, the notion of Big Data pipeline will be introduced and described with the help of a few use cases. Then we focus on the notion of Data Mining and describe the basic tasks. At last the technologies for collecting, storing and managing Big Data will be introduced and described.

10.2. Topics on Big Data Ecosystem and Tools

The second part of the course is focused on presenting the ecosystem of tools that can be used in Big Data solutions. Due to a massive interest in this area more and more tools are created or adopted so the systematisation and selection of vital ones are essential.

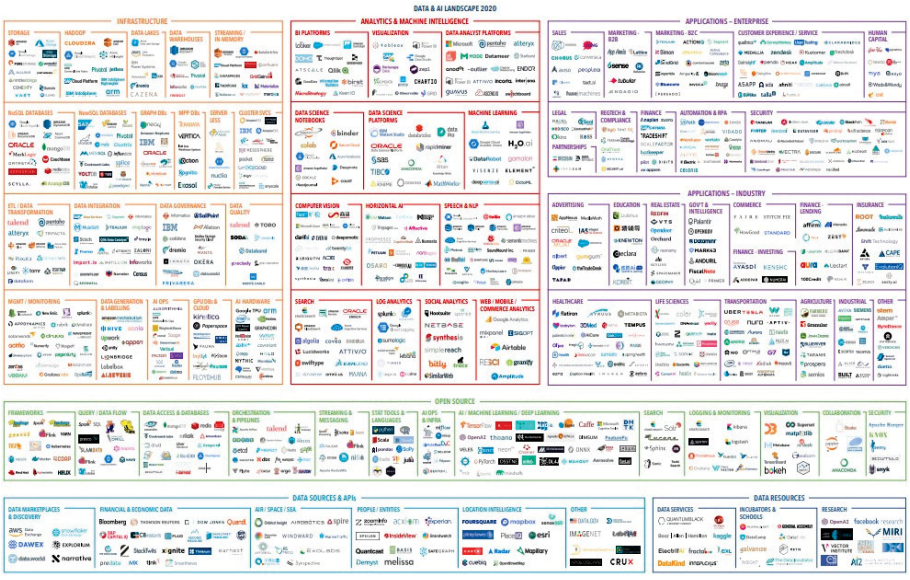

Figure 6. Technologies map by Mat Trucker [10]

The topics in following units will present to the participants the opportunities to get acquainted with Big Data tools and Hadoop ecosystem, which is a common bridge to other projects and solutions.

Section 10.2.1. Hadoop Ecosystem

Unit Outcomes

At the end of this unit the participant will have:

- Ability to process large volumes of data using hierarchical storage, hashing and filtering,
- Ability to select appropriate sampling and filtering method for given Big Data analysed case,

- Effectively use variety of data analytics techniques (Machine Learning, Data Mining, Prescriptive and Predictive Analytics),
- Using wide range of Big Data analytics platforms,
- Ability to tackle with concurrency / parallelism problems of Big Data scale.

Unit Activities

Lectures

Two lectures are developed to cover the issues related to Apache Hadoop ecosystem, which allows to process large volumes of data. The first lesson is divided into theoretical part and practical one. At the beginning, it presents the distributed file system and MapReduce concept, that allows to perform task on it. The ecosystem is presented with support application. The presentation is followed by a tutorial on installing and using basics elements of HDFS file system and MapReduce operations. The lesson is finalized by video tutorial on installation of Hadoop ecosystem.
Second lesson also follows concept of first lesson. The theoretical part presents drawbacks of MapReduce and need for using more efficient tools like Apache Spark. The installation of this tool, creating a context and basic operations was presented. Then, it is followed by comparison of Mapreduce and Apache solutions. Finally, basic functionality of this tool and Kafka tool was presented as real time and batch processing example. The presentation for each case is constructed as the sequence of various examples presenting new functionality. The main didactic goal of the lectures is to form an indicative basis for the subsequent assimilation of educational material by students. It performs scientific and educational functions, introduces the student to the world of Big Data tools and allows the lecturer to create with student environment for further training.

Teacher practical demonstration

The basic of the Big Data concept is the ability to process data with significant volume. The specific teacher training situations shows student, how to create environment based on docker. The stress is put in here to accurately present the data manipulation and describe how to explore Big Data as software practicing with an instructor. The role of "instructor" is played by the teacher, who becomes both the "organizer and participant in joint activities" to master new competences. Being in constant interaction with the students, the teacher creates such an "educational environment" in which the students themselves discover, acquire and construct their competence, show personal initiatives.

Practical tasks in a team

Within this Unit the group of students' works in teams, while working with the two environments. First environment is the instance of Hadoop with IntelliJ as the working IDE. Second one is Apache Spark library manage using Python language. The tasks allow student to enable working in this environment from scratch, while teachers are monitoring students managing their efforts and provide support to their' independent, team work, which involves:

- setting up an environment,
- configuring environment • running simple examples.
- get familiar with tools.

The command form of organization of training should fulfill three functions: integrative, communicative, managerial described above.

Q&A Sessions

In this Unit two major topics are covered in theory and practice. The questions start from HDFS structure. In here, it is essential to check the engagement of particular task member to verify, if the concept of storing Big Data is understood. Next step is to verify the practical tasks. Groups should manage the problem; however specific configuration should be verified by all members. If HDFS is familiarized, the next step is to perform knowledge check of particular tasks like map or reduce operations. Again, the theoretical knowledge should be confirmed by simple examples created based on questions. The questions concerning separate tool can be passed in direct communication with teachers or as an assignment in groups.

Unit Content

This unit introduces Apache Hadoop ecosystem and its components: Hadoop/MapReduce, HDFS, Yarn, HBase, Hive, Pig, and presents their main features and application usage. First lecture introduces Apache Hadoop ecosystem competence building which is considering introduction and its advance elements, which will allow to build students' skills and knowledge in this area.

In particular the following elements are covered:

- Apache Hadoop platform and stack of Bog Data technologies,
- Hadoop/MapReduce programming framework,
- Hadoop Distributed File System (HDFS),
- Yarn – distributed resource manager,
- HBase - NoSQL database,
- Hive – distributed warehouse built on HDFS,
- Pig – script framework for distributed data management and applications.

While going through resources of this unit the learners should be able to: get acquainted with the Apache Hadoop software ecosystem, understand principles, features and functions of Hadoop components and technologies: HDFS, Yarn, HBase, Hive, Pig and use Hadoop stack of technologies in developing Bog Data applications.

The practical task covers the configuration of Hadoop with Yarn environment. Based on it the five simple data processing examples are introduced to verify concept in practice. The student is able to count words, manage temperature readings, find relations between two consecutive words, index management and top citation example. All examples are presented as java code.

At this stage of unit the limitations of this approach are presented as three layers: efficiency (e.g. high communication cost), programming model (e.g., hard to implement everything as a MR program) and Real-time processing (e.g., a MR job requires to scan the entire input). Thus, the Apache Spark tool as remedy is presented, which process data in memory. The Spark History and its evolution is presented. Finally, RDD model processing in Spark is presented in contrast to MapReduce solution.

The next evolution in Big Data technology is the approach proposed by Apache Spark, where data are processed by data frames. The additional tools are introduced to manage to process data in this way – as SQL operations. Finally, a streaming issue

was addressed. The theory is backed by the examples in java, which will be tested in training session.
The exercises block start with the teacher demonstration. The entire environment is created using one procedure via docked image. The docked installation and configuration in explained. Using demonstration step by step all basic map reduce operations are presented. The students in group can prepare their environments. They covers the Hadoop and Apache Spark installation. As main IDE the IntelliJ is proposed.

Formative Assessment
The offered methodology requires the assessment at two levels. At first level teacher has to provide the students with lesson material, knowledge and then to verify knowledge concerning large data storage and basic processing (in parallel) in context of available tools for each task. In here the student's ability to select right tool should be verified. Then application level is verified. The student is able to run script and understand its meaning.
The unit grade is formed as a multi-component, which covers:
- knowledge of particular tool with its application case measured by the level of activity of student, while performing tasks,
- the installed environment able to run the examples pointed by teacher.

Lectures
The lectures knowledge assessment is performed by MCQs quiz. Multiple questions are organized in from that covers the abilities presented in unit outcomes. It is worth to note that questions are organized as sequence that shows the progress of trainee in lesson. This gives additional feedback to teacher at which steep the errors start to appear.

Teacher demonstration of Hadoop ecosystem tools
The presentation of basic Big Data tools ecosystem use case given by the teacher can be verified during test and consultations. It is considered satisfactory if:
- most of the tools are known by students on knowledge level;
- students can come with conclusions, which tool should be used in specific task; task can differ from batch computing to real time processing;
- adequate analytical process to select tool was presented by student.

Practical tasks in a team
At this stage student should be familiar with separate tools in Hadoop ecosystem, thus it is vital to verify their concepts in practice. The tasks should be made in teams (more than two) to take under consideration cooperation soft skills. The initial task should be followed by in team closed discussion. During team work the progress should be monitored and reported to a trainer. It is preferred that groups share their progress to other teams as public operational evaluation to stimulate the **competition**.
The installation of instances and particular operations both for MapReduce and Apache Spark should be done with understanding of basic concept, which allows to provide simple modifications to a task. **Teacher**, based on examples, can recommend changing of input file, data format or type of operation. The student must understand not only particular tool but an operation that are conducted on separate dataset.

Finally, the teacher should asses the final result based on the installation and configuration process, which has also educational effect. The estimate should have clear estimates to be considered by student as fair and just.

During practical tasks teacher have to under consideration especially development of:

- **communication skills.** Group training fits naturally into the teaching of interpersonal skills. In today's World given tasks require dialogues between clients, management and other team participants. In this context the Teacher takes a role of client and roles within system can be divided within a group. It is recommended to arbitrary select a role (force students from comport zone and at the same time allow to discover its pros and cons) and allows students to divide roles by themselves. The improvement of skills between task should be evaluated and not its level only;
- **skills of critical thinking and logical thinking** should be specially treated. Students in real world will meet with new challenges and problems on daily basis. In case of data scientist, the data type can change, its characteristic can change with client's structure changes and adaptation is needed. On the other hand, the process should be constantly improved in case of performance or complexity. Thus, the task should be presented as some iteration of solutions that can always be improved or changed. This kind of approach should be stressed in grade evaluation as well.
- **risk taking and its assessment**, which is crucial I process of decision making. The data scientist should not fear to explore and find new ways of data processing. However, in case of constant product improvements one can lost its original character. Thus, in case of improving a solution to big extend, the student should be able to estimate its rate of success and what additional value it can bring to an end user. This element should also be evaluated during tasks.

Q&A Sessions

The Q&A session is especially useful, to verify the ability of student to express its ideas and defending them. The evaluation is given in context to knowledge, skills but also one's ability to share them with others. The teacher should encourage to meaningful discussion or short impromptu speech, which should include:

1. The knowledge presentation (correctness of sentences, readiness, reasoning, etc.).
2. Drawing conclusions based on it in consideration to given topic.
3. Providing arguments for ask questions or counterproposals.
4. Demonstrate the ability of logical thinking;
5. Providing alternatives if previous one was neglected.
6. Summarizing speech with final proposition.

Resources

Questions for discussion: 1
Quiz: 1 with 30 MCQs with 4-5 answers/distractors each to assess 2 competencies
Presentations: 2
Demonstrations: 2
Learning Video: 2

Lecture notes: 2
External URLs

Section 10.2.2 Databases for Big Data
Unit Outcomes
- Ability to process large volumes of data using hierarchical storage, hashing and filtering;
- Design, build, operate relational and nonrelational databases (SQL and NoSQL);
- Using wide range of Big Data analytics platforms.

Unit Activities
Lectures
Four lectures are developed to cover the various databases types, which are used in Big Data solutions to store data. The databases are categorized into four classes: documentoriented database, key value database, column-oriented database and graph-oriented database. Each of these databases are a key for data storing and processing in various applications and Big Data pipelines. Each of it is thoughtfully described in each lecture.
The first lecture is partially dedicated to comparison of particular databases pros and cons. In this section student should get familiar with particular database type and types of data, which can be stored using its structure. The teacher should explain not only concepts but also main difference in each approach. The section is followed by document type database. The teacher is using noSQL database type to present its basic operations, its structure and show its application cases. The cases are especially vital, because they give needed background for understanding its application basis. Several use cases are presented for this task.
Second lecture are dedicated to key-value database on example of Redis instance. The lesson shows the advantages of this type of application like ability to holds database entirely in the memory. Finally, other features of Redis like transactions or kays with limited time-to-live is introduced. The main role of teacher is to stress a role of a key-value database, which uses a simple key-value method to store data. This database contains a simple string (the key) that is always unique and an arbitrary large data field (the value). It is easy to design and implement. The summary stressed the useability of instance for particular cases.
Third lecture is dedicated to column-oriented database. The one of well-known instance of this database is Cassandra database, which is designed to handle large amounts of data across many commodity servers. Unlike a table in a relational database, different rows in the same table (column family) do not have to share the same set of columns. The teacher in this lesson should show the similarity of this database to key-value model, but at the same type to stress the fact that it covers two levels of nesting. At the first level, the value of a record is in turn a sequence of key-value pairs. These nested key-value pairs are called columns, where key is the name of the column. Another important issue to note is a comparison of Casandra database and classical relation database. The application of Casandra database can be found in many Big Data solutions and is used by companies as CERN, Comcast, eBay, GitHub, Netflix or Reddit. The many companies are storing data (more than 1500) using this dataset. At the same time students should get familiar with basics of Java programming, Casandra model and Select statement dedicated of it..

Fourth lecture covers a graph database, which stores nodes and relationships instead of tables, or documents. Students should understand idea of data and be able to visualise them like a sketch of ideas on a whiteboard. The advantage of data stored in this model is lack of restrictions, which allows a very flexible way of thinking about and using it. The teacher again should put stress to differentiate the model to relational databases, where JOIN operations is time and computational consuming operation. In case of many relations classical database is not very efficient in case of large datasets. Thus, a graph database is used where no join operations are needed. Relationships are stored natively alongside the data elements (the nodes) in a much more flexible format. Everything about the system is optimized for traversing through data quickly; millions of connections per second, per core which is suitable for Big Data case. As the tool example of graph database, the Neo4j is proposed. The lecture allows to gain knowledge on the basic concept of graph database with its operation set.
The main didactic goal of the lectures in this unit is to form an indicative basis for the subsequent assimilation of educational material by students. It performs scientific and educational functions, introduces the student to the world of Databases and allows the lecturer to create with student instances to manage Big Data sets.

Teacher practical demonstration
The essence of Big Data is an ability to store them and to process them. In previous unit basics instance on Apache ecosystem was presented. In this unit the teacher is presenting in practice the instances of described in lectures database types. Teacher is using a docker to decrease the time needed to prepare the working instance of a database. The database, after configuration, is used to present the data usage and basic operations, which can be perform on it. Similar example is presented using Kubernetes. Finally, the simple frontend basing on database backend is presented. The role of "instructor" is to guide the students through the configuration and presentation process. Being in constant interaction with the students, the teacher creates such an "educational environment" in which the students themselves discover, acquire and construct their competence.

Student practical exercises
Based on the teacher presentation the students are ask to develop their competences by following the exercises. The exercises teach basic ability to use four database types used commonly in Big Data solutions. Now the teacher is introducing the series of operations and the student should follow the pattern. The teacher can increase the attractiveness of an exercises by changing to some extend the objectives. At the end of this task each student should be able to insert, process and retrieve data from various databases. Additionally, using particular examples the student gain intuition on type of data and type of database which is most suitable for its processing.

Practical tasks in a team
Within this Unit the group of students' works in teams, while working with four types of databases. The Mongodb, Redis, Casandra and Neo4j instances are considered. First the group should focus in task on type of data, which is provided, and agree to type of database instance. The brainstorm on data type, based on data structure and relation should be take under consideration. In second stage the structure of data

is created and data is store in system. Finally, the required information should be provided using set of operations. The solution should be presented and groups are elaborations on other solutions while trying to defend their implementation. The vital for this task is ability to defend a solution and admit the advantages of other solutions which can provide new better iteration in future. The command form of organization of training should fulfill three functions: integrative, communicative, managerial described above.

Every lecture is enhanced by use-case example, which add additional level of presentation. Working on those cases teachers can get to know their students better by organizing students' work in teams. This helps teachers to encourage students' independent, team work, which includes:

- the development of student motivations;
- establishing objectives and goals;
- knowledge and experience transmission;
- administrative activity;
- the coordination of student engagement;
- command of the educational process.

Q&A Sessions

In this Unit four different database implementations are considered in theory and practice. The questions are formulated in this way that they presents a particular case study e.g.

> *"If a large commercial organization decides to use MongoDB and collect and process information from its activities, how should it organize its documentary sources and document flow?"*

Therefore, in this unit, the essential for teacher is to check the engagement of particular task member and ability to tackle with particular case. The both critical thinking and creative finking should be used to find right solution. Each group member should express its proposition and try to defend it. The consensus has to be made. Then a specific configuration should be provided and verified by all members. Again, the theoretical knowledge should be confirmed by simple examples created based on questions. The questions concerning separate tool can be passed in direct communication with teachers or as an assignment in groups.

Unit Content

The unit covers the database instances, which can be used, while constructing Big Data pipelines. The Mongodb, Redis, Casandra and NEO4j databases was selected as implementation of document oriented, key value, column oriented and graph database. Each of these architectures are briefly described by showing their pross and cons. The overview shows Table 1 and Fig. 7.

Table 1. Types of non-relational databases and the features associated with them

Type	Performance	Scalability	Flexibility	Complexity
Key-value store	High	High	High	High
Column store	High	High	Moderate	Low
Document store	High	Variable to high	High	Low
Graph-based	Variable	Variable	High	High

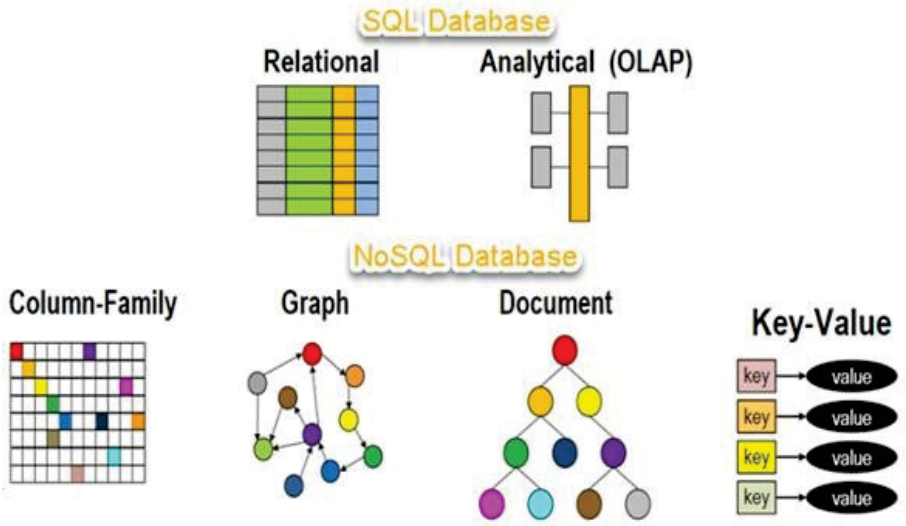

Figure 7. Types of NoSQL Databases

The approach was presented, where less ETL operation (present in MapReduce) is required and more relations and specific documents are processed like: XML or JSON. The unit also introduce ability of presented solution to tackle change over time or ability to scale horizontally of NoSQL approach. The issue of multiple structures and vendor selection was presented. Then database instances were introduced.
First one is MongoDB, which is a kind of non-relational database (NoSQL) and it is opensource. It's developed like a highly available, scalable, and fault-tolerant document-oriented solution. Instead of storing information in tables, as in traditional relational databases. The main features were presented as:

- Documents can contain many different key-value pairs, or key-array pairs, or even nested documents,
- Offers querying, indexing, and aggregation,
- Built in high availability, horizontal scaling, and geographic distribution,
- Libraries for many programming languages (C, Go, PHP, Ruby, etc.),
- Supported on Windows, Linux, and macOS,
- Hosted solution is also available by MongoDB Atlas.

The application in real cases was also presented, together with the basic installations and operation types:

- Create structure as collections,
- Process documents (insert, update, find, aggregate and drop).

Next database instance is Redis as key-value representative. Redis provides data structures such as strings, hashes, lists, sets, sorted sets with range queries, bitmaps, hyperlogs, geospatial indexes, and streams. Redis has built-in replication, Lua scripting, LRU eviction, transactions, and different levels of on-disk persistence, and provides high availability via Redis Sentinel and automatic partitioning with Redis

Cluster. To achieve top performance, Redis works with an in-memory dataset. Redis characterise by:
- perform about 110000 SETs per second, about 81000 GETs per second.
- Supports rich data types,
- Operations are atomic and can be used in a number of use cases such as caching, messaging-queues (Redis natively supports Publish/Subscribe), any short-lived data in your application, such as web application sessions, web page hit counts, etc.

The application in real cases was also presented, together with the basic installations and operation types:
- Basic GET and SET operations,
- additional operations as KEY, SELECT, FLUSHBD or INCR.

Next instance described is Cassandra, which Data Model is built with a basic key-value model, but with two levels of nesting. At the first level, the value of a record is in turn a sequence of key-value pairs. These nested key-value pairs are called columns, where key is the name of the column.
It is worth to note that Cassandra is in use at Constant Contact, CERN, Comcast, eBay, GitHub, GoDaddy, Hulu, Instagram, Intuit, Netflix, Reddit, The Weather Channel, and over 1500 more companies that have large, active data sets.
The basic features of Casandra are presented e.g.: elastic scalability, fast linear-scale performance or transaction support. Then Casandra architecture and advantage was presented. The students can get familiar with particular layers as cluster, node, column family or row. At this stage the difference between relation and Casandra database was stressed. Finally, datatype used by this database was presented.

The application in real cases was also presented, together with the basic installations and operation types:
- create key space, use catalog, create table,
- additional data processing: SELECT or UPDATE.

Last database instance described is Neo4j Graph Database, which stores all of its data in Nodes and Relationships. It stores its data in terms of Graphs in its native format. Neo4j uses Native GPE (Graph Processing Engine) to work with its Native graph storage format. The main building blocks of Graph DB Data Model are: nodes, relationships, properties and labels. The unit explain the value of each element and introduce operations to work with it using clauses of Neo4j Cypher Query Language.

As in previous cases. The knowledge is backed up by use case example. In total five use case scenarios are presented concerning fraud detection, real-time recommendation engine, master data management, network and IT operators and Identity & access management.

Formative Assessment
The methodology offered require, as in previous unit, the assessment at two levels. At first level teacher has to provide the students with lesson material, with knowledge and then verify knowledge concerning ability to process various data in parallel using

various database instances in context of available tools for each task. In here the student's ability to select right database should be verified. Then application level is verified. The student is able to instantiate the database and preform basic operation on it like creating dataset, create items or relations. Finally, obtaining required information. The unit grade is formed as a multi-component, which covers:

- knowledge of all databases, with its ability to use for research case measured by the level of activity of student, while performing tasks,
- the instantiate database with dataset and script to manage data based on task appointed by the teacher.

Lectures

The lectures knowledge assessment is performed by MCQs quiz. Multiple questions are organized in form that covers the abilities presented in unit outcomes. It is worth to note that questions are organized as sequence that shows the progress of trainee in lesson. This gives additional feedback to teacher at which database instance is strong side or weak side of particular student.

Teacher demonstration of various databases

The presentation of four various database instances backed up with use case given by the teacher can be verified during test and consultations. It is considered satisfactory if:

- document, key-value, column based and graph-based types and their pros and cons are known by students on knowledge level;
- students can come with conclusions, which database should be used in specific task; task can differ from data type provided and environment in which it should operate;
- using basic operations for each database by students cause no problem;

Practical tasks in a team

At this stage student should be familiar with each database type, thus it is vital to verify their instances in practice. The tasks should be made in teams (more than two) to take under consideration cooperation (soft skills). The initial task should be followed by in team closed discussion. The progress of each team should be monitored and the results of each stage should be reported by particular team. It is preferred, that groups share their progress to other teams, as public operational evaluation, to stimulate the competition.

Creating instances of database, proposing database scheme and list of operations should be done with understanding of basic concept. Teacher, based on examples, can recommend changing of data scheme, database type or operation sequence. The student must understand not only particular database but an operation type that can be conducted on separate dataset. Finally, the teacher should asses the final result based on process and not only a final result. The progress during the instance creation and creating commands script, which has also educational effect, should be accounted in final assessment. The estimate should have clear form to be considered by student as fair and just.

During practical tasks teacher have to under consideration especially development of **communication skills, skills of critical thinking and logical thinking and risk taking and its assessment**. This concept was extended in previous unit.

Q&A Sessions

The Q&A session is particularly beneficial in determining a student's capacity to explain and defend his or her thoughts. Knowledge, talents, and the capacity to share them with others are all considered in the evaluation. The teacher should urge students to engage in meaningful discussion or give a short spontaneous speech that includes the following points:

1. The knowledge presentation (correctness of sentences, readiness, reasoning, etc.).
2. Drawing conclusions based on it in consideration to given topic.
3. Providing arguments for ask questions or counterproposals.
4. Demonstrate the ability of logical thinking.
5. Providing alternatives if previous one was neglected.
6. Summarizing speech with final proposition.

Resources

Questions for discussion: 1
Quiz: 4 with 10 MCQs with 4-5 answers/distractors each to assess 2 competencies
Presentations: 4
Demonstrations:4
Learning Video: 2
Lecture notes: 4
External URLs

Section 10.2.3. Big Data Analytics tools

Unit Outcomes

At the end of this unit the participant will have:

- Ability to process large volumes of data using hierarchical storage, hashing and filtering,
- Ability to select the efficient algorithm to Big Data, which takes under consideration its scale,
- Ability to model, analyze, and evaluate organization's business processes,
- Ability to select appropriate sampling and filtering method for given Big Data analyzed case,
- Effectively use variety of data analytics techniques (Machine Learning, Data Mining, Prescriptive and Predictive Analytics),
- Apply quantitative techniques (statistics, time series analysis, optimization, and prediction),
- Using wide range of Big Data analytics platforms.

Unit Activities

Lectures

Four lectures are design in this unit to present Orange and Tableau tool, which can be used for various analytics tasks. The first lecture aims to reveal the big picture of the Knowledge Discovery Paradigm based on the Big Data domain, interdisciplinary links between fields, actors, and processes involved in Analytics, and the potential applications, impact, and importance on the digital business transformation, Industry 4.0 and Society 5.0. Due to scope of topics, it is narrowed to terms analysis and analytics, focusing attention on the primary types of analytics and the stages in the life-cycle. The lecture is an introductory one and place a background for practical usage of a tool. Based on this lecture teacher should verify the overall knowledge of students on analytics and its place in Big Data pipeline.

Second lecture is dedicated to data quality, which must be assured to achieve accurate results applicable to real-world use cases. Exploratory data analysis is an essential part of that. On the other hand, the most advanced results applicable to intelligent systems can be achieved through the machine learning approach. The lecture provides a basic but comprehensive introduction to EDA thought summary statistics, visualization techniques, and terminology related to Data Analytics. Machine learning methods are discussed as a request for advanced analytics. It is worth to note that the lecture is not mandatory for learners with basic knowledge in Data Mining, such as methods, algorithms, and tasks, data types, data measurements, data distribution and/or knowledge of Statistics. The teacher should focus attention of student on data analytics terminology, process of data analysis (summary statistics and visualization techniques), techniques, which assure data quality and known machine learning methods and evaluation metrics.

Third lecture described ecosystems at which dealing with the whole Big Data life-cycle, including data acquisition, analysis, data searching, storing, data sharing, data privacy and security, data visualization, scalability, orchestration, integration, interoperability, and management, is simplified task. Lecture described IT giant develop ecosystems and ecosystems of ecosystems or product families to support all necessary methods, technologies, and techniques in the Big Data processing. In this context lecture aims to reveal Big Data Technologies Landscape and its segmentation and some of the popular analytics solutions outlining their purpose, features, benefits, limitations, and training provided. Student at the end of lecture should have knowledge on segmentation of Big Data Technologies and know popular Big Data Analytics Tools.

Final lecture is dedicated to other tools, which according to various rating providers are considered as most popular tools for analytics. They are ***IBM Watson, KNIME, Orange 3*** *and* ***Tableau*** *platform.* The lecture provides a basic introduction to stated tools. The Orange and Tableau tool was selected as example platform in this unit. Thus, knowing their purpose, features, limitations is an advantage should be priority for student in this part of a Unit.

The main didactic goal of the lectures is to form an indicative basis for the subsequent assimilation of educational material by students. It performs scientific and educational functions, introduces the student vast World to analytics tools available in Big Data environment. The knowledge is essential for next practical part of Unit.

Demonstration of practical application

The analytics platform (Orange, Tableau) is presenting in four presentations two for each instances. First part considers installation and familiarization with platform and second is dedicated to data exploration. Finally, the basic classification, regression and clustering is presented step by step. The Tableau presentation follows similar pattern, however in second part forecasting, clustering, dashboard and storyline examples was introduced. The students follow the presentation, while working on two platforms. First environment is the Orange 3 instance working as a tool to analyse data. The presentation is followed by Tableau environment, where the other operations are researched. The task allows student to enable working in multiple environments. The task is considered as fulfilled if student can:

- setting up a platform,
- get familiar with tools,
- perform summary statistic, analysis, classification or prediction.

Teacher practical demonstration

The demonstration illustrates the analytics process using both Orange and Tableau platform. The specific teacher training situations shows, how to use the platform in practice. The process illustrates data import, using additional add-ons (geolocation) and using it as pipeline of data transition. In demonstration the Covid and production data was used. The role of "instructor" is played by the teacher, who becomes both the "organizer and participant in joint activities" to master new competences. Being in constant interaction with the students, the teacher creates such an "educational environment" in which the students themselves discover, acquire and construct their competence, show personal initiatives.

Practical tasks in a team

Within this Unit the group of students' works in teams, while working with two platform. First group of tasks is theoretical and allows student to categorize the various machine learning approach which exist in Orange and Tableau. Then, following the video instruction, the group should prepare data for analysis. The tasks allow student to enable working environment from scratch, while teachers are monitoring students manage get to know the students better and support their' independent, team work, which involves:

- performing prediction (multiclass, binary),
- regression,
- cluster analysis • image processing.

The command form of organization of training should fulfill three functions: integrative, communicative, managerial described above.

Q&A Sessions

In this Unit three major topics are covered. First theoretical allows student to verify their knowledge concerning machine learning methods. The topics covers supervised/unsupervised learning, dimensionality reduction or application in multiple domains. Last two topics covers application of Orange and Tableau in practical situation e.g., outliers finding. Groups should propose their solutions. Again, the theoretical knowledge should be confirmed by simple examples created

based on questions. The questions concerning separate tool can be passed in direct communication with teachers or as an assignment in groups.

Unit Content

The Unit is focused on general knowledge of analytics and its application using Orange and Tableau tool. The unit starts with introduction to the learning objectives with Analytics multi-disciplinary nature and role of diversification in that process. Next topic shows the difference between data analysis and analytics. It is stated that data analysis is a process of inspecting, cleansing, transforming, and modeling data to discover useful information, suggesting conclusions, and supporting decision-making, while data analysis has multiple facets and approaches, encompassing diverse techniques under various names in different business, science, and social science domains. Following the topic, the descriptive, diagnostic, predictive and prescriptive analytics was presented. The big focus was also put to present the lifecycle of data analytics. The several approaches were described i.e., IBM or NIST organization. Finally, five mayor steps were identified and thoughtfully described i.e.:

1. Capture: gathering and storing data, typically in its original form (i.e., raw data),
2. Preparation: processes that convert raw data into clean, organized information,
3. Analysis: techniques producing synthesized knowledge from organized information,
4. Visualization: presentation of data or analytic results in a way that communicates to others,
5. Action: processes using the synthesized knowledge to generate value.

The analysis was compared with the various sectors: government and commercial (Google, LinkedIn of Netflix).

The major part of unit is focus on exploratory data analysis. In this context machine learning approaches, techniques and algorithms are described. The main categories like classification, regression, cluster analysis, and association analysis were introduced. The basic taxonomy and summary of algorithms was presented as in exemplary Table 2.

Table 2. Basic taxonomy and summary of algorithms

Types	Name	Description	Advantages	Disadvantages
Linear	***Linear Regression***	The 'best fit' line through all data points. Predictions are numerical (regression)	Easy to understand – clearly 'view' what the biggest drivers of the model are.	Sometimes too simple to capture complex relationships between variables. A tendency for the model 'overfit.'
	Logistic Regression	The adoption of linear regression to problems of classification (e.g., yes/no questions, groups, etc.)	Also, easy to understand.	Sometimes too simple to capture complex relationships between variables. A tendency for the model 'overfit.'
Tree-based	***Decision Tree***	A graph that uses a branching method to match all possible outcomes of a decision.	Easy to understand and implement	It is often not used for prediction because it's also often too simple and not powerful enough for complex data.
	Random Forest	Takes the average of many decision trees, each of which is made with a sample of data. Each tree is weaker than a full decision tree, but we get better overall performance by combining them.	A sort of 'wisdom of the crowd'. It tends to result in very high-quality models. Fast to train	It can be slow to output predictions relative to other algorithms. Not easy to understand predictions.
	Gradient Boosting	It uses even weaker decision trees increasingly focused on 'hard' examples.	High-performing	A small change in the feature set or training set can create radical changes in the model. Not easy to understand.
Instance-based	***Support Vector Machines (SVM)***	It uses a kernels mechanism, which calculates the distance between two observations. Then it finds a decision boundary that maximizes the distance	Can model non-linear decision boundaries, & there are many kernels to choose from. Fairly robust	Memory intensive, trickier to tune due to the importance of picking the correct kernel. Don't scale well to larger datasets.

		between the closest members of separate classes. SVM with a linear kernel is similar to logistic regression.	against overfitting, especially in high-dimensional space. Easy to interpret results.	
Bayesian	***Naïve Bayes***	It is based on conditional probability and counting. The model is a probability table that gets updated through training data.	Suitable for a relatively small training set. Simple and straightforward to use. Deal with some noisy and missing data. Handles multiple classes. Easily obtain the probability for a prediction	Prone to bias when increasing the number of training sets. Assumes all features are independent and equally important, which is unlikely in realworld cases. Sensitive to how the input data is prepared.
Clustering	***K-means***	It makes clusters based on geometric distances (i.e., distance on a coordinate plane). The clusters are grouped around centroids, causing them to be globular and have similar sizes.	Relatively efficient Can process large data sets.	Applicable only when a mean is defined. Not applicable for categorical data. Unable to handle noisy data. Not suitable to discover clusters with non-convex shapes
	Hierarchical Clustering	"Bottom-up" (agglomerative) approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. "Topdown"(divisive): observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.	The main advantage of hierarchical clustering is that the clusters are not assumed to be globular. In addition, it scales well to larger datasets.	Much like K-Means, the user must choose the number of clusters (i.e., the hierarchy level to "keep" after the algorithm completes).
Neural Networks	***Neural Networks***	Mimics the behaviour of the brain. Neural networks (NN) are	Can handle highly complex tasks – no other	Very slow to train because it has so many layers.

		interconnected neurons that pass messages to each other. Deep Learning uses several layers of NN put one after the other.	algorithm comes close to image recognition.	Require much power. Almost impossible to understand predictions.

The sections are also dedicated to:

- scales at which the data can be evaluated,
- Summary statistics and their measures,
- Data exploration using plots,
- Importance in ensuring data quality.

The next section is dedicated to categorization of tools available in Big Data ecosystem. With the First Mark platform, Matt Turck visualized the seventh annual Big Data & AI Landscape for 2020 available at https://mattturck.com/data2020/. As the visualization is large with many tools posted and a particular interest in Analytics, a segment from the whole landscape is presented. The roadmap allows to catch general understanding on ecosystems and well-rating Big Data Analytics products basis on rating platforms: Trust Radius, BI Survey, Gartner Peer Insights, Select Hub. Based on it the popular Big Data analytics tools was presented:

- **Cloudera** is unified BD platform including set of tools,
- **Oracle Analytics** Cloud ***is*** an AI-powered solution that provides robust reporting and analytics features to businesses of all sizes,
- **SAP HANA** is the in-memory database for SAP's Business Technology platform with solid data processing and analytics capabilities,
- **The Alteryx** platform is a suite of five products offering self-service statistical, predictive, and spatial data analytics to achieve enterprise, financial and industrial intelligence,
- **SAS Viya** is a cloud-based in-memory analytics engine that provides data visualization, reporting, and analytics to businesses for actionable data insights,
- **Apache Spark** is an open-source unified analytics software for distributed, rapid processing,
- ***Qlik Sense*** is a self-service data analytics software that enhances human intuition with the power of artificial intelligence to enable better data-driven business decisions,
- ***MicroStrategy*** is a data analytics platform that delivers actionable intelligence to organizations of all sizes,
- ***Azure Big Data*** from Microsoft provides cloud computing robust analytics and AI services
- Others with dedicated languages like **Python** or **R**.

Final section covers four most popular platform for data analytics They are IBM Warson, Knime, Orange 3 and Tableau. In case of all platform their purpose, features and operation principles are described. In case of IBM Watson the various produces, that are part of this platform was characterized. Finally, some remarks considering its

applications was made. The most valuable part is summary, where the student can find application study of each solution. In case of KNIME e.g., "we get a robust open-source solution with cross-platform interoperability. It integrates with a range of software, such as JS, R, Python, and Spark. With various nodes and functions, it can process large datasets with a decent level of control in each step […]".

Formative Assessment

The methodology is suitable to the level of assumed competency i.e., application. Thus, the teacher has to provide the students with lesson material, with knowledge and then verify its results concerning data analysis platforms and its application for particular data sets. In here the student's ability to select right platform and analysis algorithm should be verified. The verification considering the creation of pipeline of data process using those tools and at the same time present its meaning. The unit grade is formed as a multi-component, which covers:

- knowledge of various data analytics techniques and its application for particular domains - knowledge of machine learning methods, which can be used for summary, classification or clustering,
- ability to use Orange and Tableau tool to perform data summary/statistic, while performing tasks,
- adapt examples to requirement defined by a teacher.

Lectures

The lectures knowledge assessment is performed by MCQs quiz. Multiple questions are organized in from that covers the abilities presented in unit outcomes. It is worth to note that questions are organized in three groups which allows to better track a progress of trainee in lesson. The feedback gives a teacher information, which part of unit should be repeated.

Teacher demonstration

The presentation of usage of analytical platforms uses simple examples provided by the teacher can be verified during test and consultations. It is considered satisfactory if:

- most of the tools are known by students on knowledge level;
- students can come with conclusions, which tool should be used in specific task; task can differ depending on data type and its dimensionality,
- use adequate algorithms to process the data.

Practical tasks in a team

At this stage student should be familiar with analytical aspects, its types and tools which are used in analysis. Practical task is vital to verify learned concepts in practice. At first step two practical (self-learning) examples are provided. They aim to verify the basic actions as installing the software, familiarization with data workspace, using limited pre-processing functionality, using various visualisation techniques by provided video and text book. The challenge using the instructor dataset is then presented. To take into account cooperative soft skills, the exercises should be completed in groups (more than two). Following the initial task, a confidential team

discussion should be held. During teamwork, progress should be tracked and each team's progress should be reported. To encourage competitiveness, groups should share their progress with other teams as a public operational review.
The installation of Orange and Tableau instances and preforming tasks should be done with understanding a concept of data analytics and taking under consideration data quality. Teacher, based on examples, can recommend changing of input file, data format or number of attributes. The student must understand not only particular tool but an operation that are conducted on separate dataset. Finally, the teacher should asses the final result based on the logic behind the produce pipeline and not result. The estimate should have clear estimates to be considered by student as fair and just.

During practical tasks teacher have to under consideration especially development of: **communication skills, skills of critical thinking and logical thinking and risk taking and its assessment**.

Q&A Sessions
The Q&A session is very beneficial for determining a student's ability to express and defend their ideas. The assessment is based on one's knowledge, talents, and capacity to share them with others. The teacher should encourage significant discussion or a brief impromptu speech that includes the following points:

1. The display of information (correctness of sentences, readiness, reasoning, etc.).
2. Drawing judgments based on it in light of the subject.
3. Providing counter-arguments to queries or suggestions.
4. Display logical thinking abilities.
5. Provide alternatives if the prior one was overlooked.
6. Concluding the speech with a final proposition.

Resources
Questions for discussion: 1
Quiz: 3 with 20 MCQs with 4-5 answers/distractors each to assess 2 competencies
Presentations: 4
Demonstrations: 4
Learning Video: 3
Lecture notes: 4
External URLs

10.3. Topics on Big Data Solutions in a Real World

This is the third stage in these Topics, were the teacher will present to the participants the opportunities of Big Data techniques based on use cases from real world applications.

Section 10.3.1. Exploratory Data, Data Analysis

Unit Outcomes
At the end of this unit the participant will be able to:

- Explorate Big Data with the help of aggregating statistics,
- Explorate Big Data with the help of visualization of the data,
- Evidence relations between the attributes of Big Data,

- Use specialized data structures when coping with Big Data,
- Solve the problems of Big Data exploration for various subject areas.

Unit Activities

Lectures

Two lectures are developed to cover the issues related to exploratory data analysis when being applied to real world use cases. The lectures are focusing both various branches of use cases and application of different software tools in Big Data. The presentation for each case is constructed in the sequence of steps: question – code – answer – conclusions.

The main didactic goal of the lectures is to form an indicative basis for the subsequent assimilation of educational material by students. Being the main link of the didactic cycle of education, it performs scientific and educational functions, introduces the student to the Big Data real-world use cases with the help of the lecturer's creative laboratory.

The lectures serve as the methodological and organizational basis for all forms of training sessions, including independent ones. The methodological basis, since it introduces the student to Big Data EDA in general, gives the unit a conceptuality, and the organizational one, since all other forms of unit activities are “tied” to the lecture in one way or another, most often logically follow it, rely on it meaningfully and thematically.

Demonstration of real-world use cases

The essence of using the Big Data use cases is the use of specific training situations, descriptions of certain environment from real world, when organizing the learning process, guiding students to formulate a problem and search for options for solving it with subsequent analysis in a team. Quite accurately the essence of this method can be described as “to learn how to explore Big Data, you do not need long lectures on software engineering, you need software practicing with an instructor.”

The role of "instructor" is played by the teacher, who becomes both the "organizer and participant in joint activities" to master new competences. Being in constant interaction with the students, the teacher creates such an "educational environment" in which the students themselves discover, acquire and construct their competence, show personal initiatives.

Practical tasks in a team

The organization of students' work in teams, while working with use cases on Big Data EDA, allows teachers to get to know the students better and support students' independent, team work, which involves:

- information of motivating motives;
- setting goals and objectives;
- transfer of knowledge and experience;
- organizational activity;
- organization of interaction between students;
- control of the learning process.

The command form of organization of training can perform three specific functions: integrative, communicative, and managerial.

The integrative function lies in the fact that the goals, content, methods and means of teaching form signs of consistency, accessibility as a result of the interaction between the teacher and students.
The second distinctive function of practical tasks in teams is communicative. The activity and nature of communication between students and the teacher and between students themselves depends on the organization of communication in the training process. The team form of organizing activities requires a high level of professionalism and a culture of communication.
Creating use case related to Big Data EDA during training in the classroom provides more opportunities for interaction between participants in the communication process.
The third function is managerial. It can be considered as a means of managing the training, education and development of trainees and at the same time as preparing future specialists for management activities.

Q&A Sessions
It is another important, and perhaps the key one for the involvement of students in the creation of the activity content on Big Data EDA, is direct communication with teachers directly at the training (online or offline). Q&A session, if it is slightly modified, modernized, this form of communication can just become the simplest and most direct way for the students themselves to form the content value of the activity.

Learning Scenarios
In learning scenarios related to Big Data EDA, the students are actively involved in the process from start to finish.
The activity is opposed to traditional didactic teaching, where information is presented directly, or there is a standardized methodology for acquiring knowledge. The process of cognition is controlled by a teacher as an intermediary. Students must identify and explore problems and questions in order to expand their knowledge or find solutions. Discovery learning includes problem-based learning, and is typically based on research and small use cases, as well as academic research. Learning by discovery is very closely related to the development and practice of critical thinking.
The cognitive processes that people participate in while learning through discovery include the following:

- Asking Your Own Questions,
- Gathering evidence that helps answer the question(s),
- Explaining collected evidence,
- Linking explanations to the knowledge they came up with during the exploratory process,
- Creating arguments and justifications for why the explanation is valid.

Discovery learning includes asking questions, noticing details, checking what information has already been learned, developing methods for conducting experiments, developing tools for collecting data, collecting, analyzing and interpreting data, pointing out possible explanations, predicting future research.
Scenario-based learning is justified for the unit related to Big Data EDA because:

- the decision made at a certain moment affects how everything goes on;

- the task requires analysis and problem-solving skills; - there is no single correct solution to the problem; - difficult to provide practical experience.

Learning scenarios can be linear as well as non-linear (branching).
Branching is the choice of a sequence of actions depending on the fulfillment or nonfulfillment of a certain condition. Branching in training scenarios makes it possible to build logical chains in order to optimally solve the problem with the least possible losses.

Unit Content
This topic is considering a key feature of exploratory data analysis (EDA) when being applied to Big Data real-world use cases in medicine and biology, sociology and demographics, ecology and weather, navigation, financing and business, art and literature, science.
EDA is not a process with a strict sequence of steps. It is rather an iterative cycle of steps during which you:

- generate the questions related to your data (some ideas to check); - answer to the questions with the help of summarizing statistics, data transformation and visualization;
- generate new refined questions basing on the previous answers and so on.

Visualization is closely dealt with EDA and aims to serve for the following goals of data exploration:

- understanding the distributional characteristics of variables,
- detecting data entry issues,
- identifying outliers in the data,
- understanding relationships among variables,
- selecting suitable variables for data analysis (feature extraction).

Python offers PySpark being API for Apache Spark, an open source, distributed computing framework and set of libraries for real-time, Big Data processing. If you're already familiar with Python and libraries such as Pandas, then PySpark will be a good language to learn to create more scalable analyses and pipelines.
R offers special packages and data types when working with Big Data. For example data.table (from package data.table) is high-performance analogue of data.frame. The working with data.table is conceptually similar with SQL:
The sparklyr package gives us an R interface to Apache Spark and a complete dplyr functionalities. Apache Spark can also be accessed with the help of the sparkR package provided by Apache.
Use cases for EDA include example from various branches of real-world application, namely

- medicine and biology,
- sociology and demographics,
- ecology and weather,
- navigation,
- financing and business,
- art and literature,
- science.

The Table 3 presenting the general description of use cases.

Table 3. General description of use cases

	Use case (link to dataset)	**branch**	**types of data used**
1	COVID-19 (numeric, labels) (*https://github.com/covid19datahub/R*)	public health	numeric, characters, and labels
2	Historical Daily Weather Data 2020 (*https://www.kaggle.com/vishalvjoseph/weather-dataset-for-covid19-predictions*)	weather	numeric, characters, and labels
3	Global Health Data Exchange (*http://ghdx.healthdata.org/*)	public health	numeric, characters, and labels
4	Sensors of air pollution (geolocational data)	ecology	numeric, characters, and labels
5	Vessels segmentation (images) (*https://www.idiap.ch/software/bob/docs/bob/bob.db.drive/stable/index.html*)	medicine	images
6	Google Street View House Number (SVHN) Dataset (images) (*https://github.com/aditya9211/SVHN-CNN*)	navigation	images
7	Physionet (signals) (physionet.org)	medicine	signal
8	Natural Language Processing (texts) and application for Stocks	literature	text
9	Stock exchange data (*https://www.kaggle.com/mattiuzc/stockexchange-data*)	finance	numeric, characters, and labels
10	Biochemical reaction study	science	signal

Formative Assessment

The methodology offered requires the assessment of not so much a set of specific knowledge as the ability of students to analyze a specific use case, make a decision, think logically, while it is best to use a multi-component method for forming the final grade, the components of which will be grades for:

- participation in a discussion and presentation, measured by the level of activity student - for prepared works.

Lectures

The assessment of lectures is organized as quiz in the form of MCQs where each MCQ is corresponding to the following competences:

- Ability to select the efficient algorithm to Big Data, which takes under consideration its scale,
- Ability to select appropriate sampling and filtering method for given Big Data analyzed case.

Demonstration of real-world use cases

The analysis of the Exploratory Data Analysis (EDA) Big Data use case given by the student during a non-public (written) presentation is considered satisfactory if:

- most of the problems in the use case have been formulated and analyzed; - carried out the maximum possible number of computing for the purpose of Big Data exploration;
- own conclusions were made based on the information about the EDA Big Data use case, which differ from the conclusions of other students;
- adequate analytical methods for information processing have been demonstrated;
- the documents drawn up in terms of meaning and content meet the requirements; - the arguments given as a result of the analysis are in accordance with the previously identified problems, the conclusions drawn, the assessments and the analytical methods used.

A serious problem in the application of the use case method in EDA Big Data studying is its role in shaping the assessment of student knowledge in the **entire course**. There are **three** possible solutions to this problem.

The **first** option is based on the assumption that the EDA Big Data use case reflects the key provisions of the system of knowledge and skills that the student must master, so the grade received by the student in the case can act as his grade in the discipline.

The **second** option proceeds from the position that the EDA Big Data use case method is not a universal method for obtaining, and even more so for assessing a student's knowledge, therefore, it needs to be supplemented by other methods, which are: oral or written exam, written work, test. In this case, the assessment received by the student from the analysis of the use case is given a certain quota of points.

The **third** option comes from an even greater commitment to other assessment methods. In this case, the user case-study method is considered as one of the many methods used in teaching this course.

Using the use case method for Big Data EDA , you can use all types of assessments: current, intermediate and final.

The **current** assessment helps to guide the discussion of the use case; an **intermediate** assessment allows you to record the progress of a student along the path of solving a use case; the **final** one sums up the student's success in case analysis and mastering the Big Data training course.

Practical tasks in a team

When evaluating the work of teams (subteams) in an open discussion, public operational evaluation of the current work of the team (subteam) can be used, which stimulates **competition**.

It should be emphasized that the evaluative creativity of the teacher should be justified. The student must understand not only the rules for analyzing the use case, but also the system of its evaluation by the teacher, the latter requires its mandatory clarification before starting work on the use case. The teacher should not forget about the **educational** effect of assessment, due not only to the openness and understandability of the assessment system for the student, but also to its fairness.

Q&A Sessions

Any word spoken by a student during Q&A Session cannot be automatically added to his asset. It is necessary to evaluate a student for meaningful activity in a discussion or public (oral) presentation, which includes the following components:

1. Speech that characterizes an attempt at a serious preliminary analysis (correctness of sentences, readiness, reasoning, etc.).
2. Drawing attention to a certain range of issues that require in-depth discussion.
3. Possession of the categorical apparatus, the desire to give definitions, to identify the content of concepts.
4. Demonstration of the ability to think logically, if the points of view expressed earlier are summed up and lead to logical conclusions.
5. Offering alternatives that were previously neglected.
6. Proposing a specific plan of action or a plan for implementing a solution.
7. Identification of essential elements that should be taken into account in the analysis of the use case.
8. Significant participation in the processing of quantitative data, computing.
9. Summing up the discussion.

Learning Scenarios
With the help of scenario-based learning, it is possible to assess:

Communication skills. Scenarios-based learning fits naturally into the teaching of interpersonal skills. Branching scenarios can effectively simulate dialogues between a data scientist and a client, a data scientist and a client's network, and between data scientists. Building dialogue simulators with a branching scenario immerses the student in a role-playing game that allows students to apply knowledge in a realistic context and receive meaningful feedback as a reaction from a virtual interlocutor.

Skills of critical thinking, logical thinking, thinking based on the principles of EDA. On the one hand, the data scientist involves the performance of routine tasks that require the same sequence of actions, on the other hand, the preparation of the data requires a deep analysis and understanding of the real situation. It is learning with the help of scenarios of real-world use cases that can provide an opportunity to gain experience in a virtual safe environment for the experimental data and subsequently implement this experience in the real-world application of Big Data EDA.
Skills for high-risk tasks. In real practice, the price of a mistake can be very high. Of course, no learning scenario can replace real experience, but at least learners will be able to make some mistakes and learn from them in a simulated environment without the risk of mistaken decisions made, data lost, or other serious consequences.

Resources
Questions for discussion: 5
Quiz: 1 with 40 MCQs with 5 answers/distractors each to assess 2 competencies
Presentations: 3
Demonstrations: 2
Learning Video: 3
Lecture notes: 2
External URLs

Section 10.3.2. Classification methods, Neural Networks
Unit Outcomes
At the end of this unit the participant will be able to

- Gather algorithms that can be used to problem-solving, the importance of expert knowledge, and being open to the views of experts .
- Big data libraries to building a classification and prediction models.
- Quantity techniques used in classification systems.

Unit Activities

Lectures

Two lectures are developed to cover the issues related to exploratory data analysis when being applied to real world use cases. The lectures are focusing both various branches of use cases and application of different software tools in Big Data. The presentation for each case is constructed in the sequence of steps: question – code – answer – conclusions.

The main didactic goal of the lectures is to form an indicative basis for the subsequent assimilation of educational material by students. Being the main link of the didactic cycle of education, it performs scientific and educational functions, introduces the student to the Big Data real-world use cases with the help of the lecturer's creative laboratory.

The lectures serve as the methodological and organizational basis for all forms of training sessions, including independent ones. The methodological basis, since it introduces the student to Big Data EDA in general, gives the unit a conceptuality, and the organizational one, since all other forms of unit activities are "tied" to the lecture in one way or another, most often logically follow it, rely on it meaningfully and thematically.

Demonstration of real-world use cases

The essence of using the Big Data use cases is the use of specific training situations, descriptions of certain environment from real world, when organizing the learning process, guiding students to formulate a problem and search for options for solving it with subsequent analysis in a team. Quite accurately the essence of this method can be described as "to learn how to explore Big Data, you do not need long lectures on software engineering, you need software practicing with an instructor."

The role of "instructor" is played by the teacher, who becomes both the "organizer and participant in joint activities" to master new competences. Being in constant interaction with the students, the teacher creates such an "educational environment" in which the students themselves discover, acquire and construct their competence, show personal initiatives.

Practical tasks in a team

The organization of students' work in teams, while working with use cases on Big Data EDA, allows teachers to get to know the students better and support students' independent, team work, which involves:

- formation of motivating motives;
- setting goals and objectives;
- transfer of knowledge and experience;
- organizational activity;
- organization of interaction between students;
- control of the learning process.

The command form of organization of training can perform three specific functions: integrative, communicative, managerial.
The **integrative** function lies in the fact that the goals, content, methods and means of teaching form signs of consistency, accessibility as a result of the interaction between the teacher and students.
The second distinctive function of practical tasks in teams is **communicative**. The activity and nature of communication between students and the teacher and between students themselves depends on the organization of communication in the training process. The team form of organizing activities requires a high level of professionalism and a culture of communication.
Creating use case related to Big Data EDA during training in the classroom provides more opportunities for interaction between participants in the communication process.
The third function is **managerial**. It can be considered as a means of managing the training, education and development of trainees and at the same time as preparing future specialists for management activities.

Q&A Sessions
It is another important, and perhaps the key one for the involvement of students in the creation of the activity content on Big Data EDA, is direct communication with teachers directly at the training (online or offline). Q&A session, if it is slightly modified, modernized, this form of communication can just become the simplest and most direct way for the students themselves to form the content value of the activity.

Learning Scenarios
In learning scenarios related to Big Data EDA, the students are actively involved in the process from start to finish.
The activity is opposed to traditional didactic teaching, where information is presented directly, or there is a standardized methodology for acquiring knowledge. The process of cognition is controlled by a teacher as an intermediary. Students must identify and explore problems and questions in order to expand their knowledge or find solutions. Discovery learning includes problem-based learning, and is typically based on research and small use cases, as well as academic research. Learning by discovery is very closely related to the development and practice of critical thinking.
The cognitive processes that people participate in while learning through discovery include the following: - Asking Your Own Questions

- Gathering evidence that helps answer the question(s),
- Explaining collected evidence,
- Linking explanations to the knowledge they came up with during the exploratory process - Creating arguments and justifications for why the explanation is valid.

Discovery learning includes asking questions, noticing details, checking what information has already been learned, developing methods for conducting experiments, developing tools for collecting data, collecting, analyzing and interpreting data, pointing out possible explanations, predicting future research.
Scenario-based learning is justified for the unit related to Big Data EDA because:

- the decision made at a certain moment affects how everything goes on;

- the task requires analysis and problem-solving skills; - there is no single correct solution to the problem; - difficult to provide practical experience.

Learning scenarios can be linear as well as non-linear (branching).
Branching is the choice of a sequence of actions depending on the fulfillment or nonfulfillment of a certain condition. Branching in training scenarios makes it possible to build logical chains in order to optimally solve the problem with the least possible losses.

Unit Content
This topic is considering a key feature of exploratory data analysis (EDA) when being applied to Big Data real-world use cases in medicine and biology, sociology and demographics, ecology and weather, navigation, financing and business, art and literature, science.
EDA is not a process with a strict sequence of steps. It is rather an iterative cycle of steps during which you:

- generate the questions related to your data (some ideas to check);
- answer to the questions with the help of summarizing statistics, data transformation and visualization;
- generate new refined questions basing on the previous answers and so on.

Visualization is closely dealt with EDA and aims to serve for the following goals of data exploration:
- understanding the distributional characteristics of variables,
- detecting data entry issues,
- identifying outliers in the data,
- understanding relationships among variables,
- selecting suitable variables for data analysis (feature extraction).

Python offers PySpark being API for Apache Spark, an open source, distributed computing framework and set of libraries for real-time, Big Data processing. If you're already familiar with Python and libraries such as Pandas, then PySpark will be a good language to learn to create more scalable analyses and pipelines.
R offers special packages and data types when working with Big Data. For example data.table (from package data.table) is high-performance analogue of data.frame. The working with data.table is conceptually similar with SQL:
The sparklyr package gives us an R interface to Apache Spark and a complete dplyr functionalities. Apache Spark can also be accessed with the help of the sparkR package provided by Apache.
Use cases for EDA include example from various branches of real-world application, namely:
- medicine and biology,
- sociology and demographics,
- ecology and weather,
- navigation,
- financing and business,
- art and literature,
- science.

Formative Assessment

The methodology offered requires the assessment of not so much a set of specific knowledge as the ability of students to analyze a specific use case, make a decision, think logically, while it is best to use a multi-component method for forming the final grade, the components of which will be grades for:

- participation in a discussion and presentation, measured by the level of activity student - for prepared works.

Lectures

The assessment of lectures is organized as quiz in the form of MCQs where each MCQ is corresponding to the following competences:

- Ability to select the efficient algorithm to Big Data, which takes under consideration its scale,
- Ability to select appropriate sampling and filtering method for given Big Data analyzed case,

Demonstration of real-world use cases

The analysis of the EDA Big Data use case given by the student during a non-public (written) presentation is considered satisfactory if:

- most of the problems in the use case have been formulated and analyzed; - carried out the maximum possible number of computing for the purpose of Big Data exploration;
- own conclusions were made based on the information about the EDA Big Data use case, which differ from the conclusions of other students;
- adequate analytical methods for information processing have been demonstrated;
- the documents drawn up in terms of meaning and content meet the requirements;
- the arguments given as a result of the analysis are in accordance with the previously identified problems, the conclusions drawn, the assessments and the analytical methods used.

A serious problem in the application of the use case method in EDA Big Data studying is its role in shaping the assessment of student knowledge in the **entire course**. There are **three** possible solutions to this problem.

The **first** option is based on the assumption that the EDA Big Data use case reflects the key provisions of the system of knowledge and skills that the student must master, so the grade received by the student in the case can act as his grade in the discipline.

The **second** option proceeds from the position that the EDA Big Data use case method is not a universal method for obtaining, and even more so for assessing a student's knowledge, therefore, it needs to be supplemented by other methods, which are: oral or written exam, written work, test. In this case, the assessment received by the student from the analysis of the use case is given a certain quota of points.

The **third** option comes from an even greater commitment to other assessment methods. In this case, the user case-study method is considered as one of the many methods used in teaching this course.

Using the use case method for Big Data EDA , you can use all types of assessments: current, intermediate and final.

The **current** assessment helps to guide the discussion of the use case; an **intermediate** assessment allows you to record the progress of a student along the path of solving a use case; the **final** one sums up the student's success in case analysis and mastering the Big Data training course.

Practical tasks in a team

When evaluating the work of teams (subteams) in an open discussion, public operational evaluation of the current work of the team (subteam) can be used, which stimulates **competition**.

It should be emphasized that the evaluative creativity of the teacher should be justified. The student must understand not only the rules for analyzing the use case, but also the system of its evaluation by the teacher, the latter requires its mandatory clarification before starting work on the use case. The teacher should not forget about the **educational** effect of assessment, due not only to the openness and understandability of the assessment system for the student, but also to its fairness.

Q&A Sessions

Any word spoken by a student during Q&A Session cannot be automatically added to his asset. It is necessary to evaluate a student for meaningful activity in a discussion or public (oral) presentation, which includes the following components:

1. Speech that characterizes an attempt at a serious preliminary analysis (correctness of sentences, readiness, reasoning, etc.).
2. Drawing attention to a certain range of issues that require in-depth discussion.
3. Possession of the categorical apparatus, the desire to give definitions, to identify the content of concepts.
4. Demonstration of the ability to think logically, if the points of view expressed earlier are summed up and lead to logical conclusions.
5. Offering alternatives that were previously neglected.
6. Proposing a specific plan of action or a plan for implementing a solution.
7. Identification of essential elements that should be taken into account in the analysis of the use case.
8. Significant participation in the processing of quantitative data, computing.
9. Summing up the discussion.

Learning Scenarios

With the help of scenario-based learning, it is possible to assess:

Communication skills. Scenarios-based learning fits naturally into the teaching of interpersonal skills. Branching scenarios can effectively simulate dialogues between a data scientist and a client, a data scientist and a client's network, and between data scientists. Building dialogue simulators with a branching scenario immerses the student in a role-playing game that allows students to apply knowledge in a realistic context and receive meaningful feedback as a reaction from a virtual interlocutor.

Skills of critical thinking, logical thinking, thinking based on the principles of EDA. On the one hand, the data scientist involves the performance of routine tasks that require the same sequence of actions, on the other hand, the preparation of the

data requires a deep analysis and understanding of the real situation. It is learning with the help of scenarios of real-world use cases that can provide an opportunity to gain experience in a virtual safe environment for the experimental data and subsequently implement this experience in the real-world application of Big Data EDA.

Skills for high-risk tasks. In real practice, the price of a mistake can be very high. Of course, no learning scenario can replace real experience, but at least learners will be able to make some mistakes and learn from them in a simulated environment without the risk of mistaken decisions made, data lost, or other serious consequences.

Resources
Questions for discussion: 5
Quiz: 1 with 40 MCQs with 5 answers/distractors each to assess 2 competencies
Presentations: 3
Demonstrations: 2
Learning Video: 3
Lecture notes: 2
External URLs

10.4. Statistical methods

The unit is the continuation of the previous units on exploratory data analysis and the analytics stage. It is focusing on applying statistical techniques for solving Big Data problems on a series of use cases from various branches. Upon completion of this unit, learners should be able to:

- select the efficient algorithm to Big Data, which takes under consideration its scale,
- effectively use variety of data analytics techniques (Machine Learning, Data Mining, Prescriptive and Predictive Analytics) from viewpoint of application to Big Data,
- apply quantitative techniques (statistics, time series analysis, optimization, and prediction) from viewpoint of Big Data in real-world use cases.

The unit is based on the same activities and the formative assessment techniques as the previous ones.

Conclusion

It takes a lot of effort, time, and resources to run a program like this. Trainers, participants (students), mentors, and business owners (managers) are the four groups that must be involved in this program, each playing a particular function.
A minimum of 12 people is recommended. This will allow working in at least three groups of four people to come up with a solution to the business challenge.
The number of the groups can vary, but they should preferably be between 4 and 5. It's also a good idea to gauge mentors' and business owners' (managers') enthusiasm, as they play an important part in the program.

The general structure is centered on students and designed to meet the demands of diverse businesses. Teachers trained students to give a framework of theories, ideas, thoughts, and concepts to build understanding through a series of workshops and work with real cases.

REFERENCES:

1. iBigWorld: Innovations for Big Data in a Real World (Erasmus+ project 2020-1-PL01-KA203-082197) documentation *https://ibigworld.ath.edu.pl/index.php/en/home-english/*
2. MANZ E., PARKER RENGA I.: Understanding how teachers guide evidence construction conversations." Science Education 101.4 (2017): 584-615.
3. MILLS G. E.: Action research: A guide for the teacher researcher. Prentice-Hall, Inc., One Lake Street, Upper Saddle River, New Jersey 07458, 2000.
4. SMITHERS A., ROBINSON P., COUGHLAN M.-D.: The good teacher training guide 2012. Centre for Education and Employment Research University of Buckingham: Buckingham, UK (2012): 27-28.
5. McCOY C, SHIH P." Teachers as producers of data analytics: A case study of a teacher-focused educational data science program. Journal of Learning Analytics 3.3 (2016): 193-214.
6. HAZZAN O., LAPIDOT T, RAGONIS N.: Guide to teaching computer science. Springer International Publishing, 2020.
7. MEERBAUM-SALANT O., ARMONI M., BEN-ARI M.: Learning computer science concepts with scratch. Proceedings of the Sixth international workshop on Computing education research. 2010.
8. HICKS S. C., IRIZARRY R.A.: A guide to teaching data science. The American Statistician 72.4 (2018): 382-391.
9. HAZZAN O., RAGONIS N, LAPIDOT T.: Data science and computer science education. Guide to Teaching Computer Science. Springer, Cham, 2020. 95-117.
10. Webpage: *http://mattturck.com/wpcontent/uploads/2020/09/2020-Data-and-AI-Landscape-Matt-Turck-at-FirstMark-v1.pdf*