Co-funded by the
Erasmus+ Programme
of the European Union

Erasmus+ SkoPS

# <D0.2: Big Data Management Course Outline and Description>

| | | | |
|---|---|---|---|
| **Project Title** | Empowering the European Workforce Development through Online/Virtual Skills Training for Digital Transformation towards Mitigating the Impact of Pandemic Situations (SkoPS) | | |
| **Project Acronym** | SkoPS | **Project Number** | 2020-1-DE01-KA226-HE-005772 |
| **Date** | 2021-11-03 | **Deliverable No.** | D0.2 |
| **Contact Person** | | **Organisation** | Petanux (PTX) |
| **Phone** | - | **E-Mail** | Mahnaz.mirhaj@petanux.com |
| **Version** | 2 | **Confidentiality level** | Public |

## Version History

| Version No. | Date | Change | Editor(s) |
|---|---|---|---|
| 1 | 13.12.2021 | Initial draft | Mahnaz Mirhaj |
| 2 | 14.12.2021 | Second Draft | Mahnaz Mirhaj |

## Contributors

| Name | Organization |
|---|---|
| **Mahdi Bohlouli** | PTX |
| **Mahnaz Mirhaj** | PTX |
| | |

## Disclaimer

This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

# Table of Contents

# 1 Introduction

In the modern world, high volumes of data are generated at a rapid pace. To manage, control, and employ this amount of data Big Data Management technology can be a great asset. Moreover, utilizing Big Data Management strategies and tools leads businesses to enhance their decision-making and improve their outcomes; therefore, learning it has become a necessity.

## 1.1 Abstract

The Big Data Management course is arranged in nine chapters aiming to help students enhance their knowledge and skills in the most updated policies and tools of Big Data Management. The course covers the technical, theoretical, and practical parts of Big Data Management. This course offers different types of activities and media to assist students to understand Big Data Management and employ it in different areas.

## 1.2 Purpose of the document

The purpose of this document is to have an outline for the Big Data Management course. In this file, the general structure of the Big Data Management course is described and determined. This file includes description, materials, activities, objective, contents, prerequisites, references, assessment methods of the Big Data Management course.

## 1.3 Relation to other deliverables

The Big Data Management course is in relation to Fundamental Machine Learning, Smart Cities and IoT, and Python programming for IoT and data science course.

# 2 Template

| Course Plan Template | | | |
|---|---|---|---|
| **Course ID and Title:** | Big Data Management | | |
| **Course Duration:** | 6 Weeks | **Course ECTS:** | |
| **Leading Organization:** | Petanux | | |
| **Course Media:** | Video, Text File | | |
| **Laboratory (Yes/No)** | Yes | | |

**Course Description:**

Big data management is one of the state-of-the-art technologies that refer to organizing, controlling, storing, and processing vast and complex data, whether the data is structured, unstructured, or semi-structured. Big Data brings insights to drive novel techniques and strategies in industrial and scientific works as well as establishing high data quality for big data analysis.

This course provides modern, approved, and most favorable Big Data Management methods, challenges, and ways of addressing them. This course prepares you to utilize Big Data Management strategies and techniques to govern big data, locate valuable information, and analyze data and results. To this aim, this course covers both the theoretical and practical parts of Big Data Management. For the theoretical part, we plan to address the basics of Big Data Management, Data Mining, Data Stream Mining, finding frequent items, and the Map-reduce technique by using PowerPoint slides mixed with some further video tutorials from the web. In addition, to educate you practically we have designed a virtual laboratory that concentrates on applications of Big Data Management, using Apache Spark tool.

This course is planned for students of Computer Science, Mathematics, and any other related fields in bachelor or master degree or those who want to learn Big Data Management technologies on a self-learning basis. To use this material you need to have a solid knowledge of Algorithm Design and be familiar with Linear Algebra and Statistics. Besides, for implementing the hands-on project while you are not expected to be an expert in coding, it is necessary to know some basics and have access to the Apache Spark environment.

Students and participants of this course are expected to learn the basics of Big Data Management, concepts and algorithms of Big Data Management, get familiar with Data Mining/ Data Stream Mining, and learn how to work with Big Data Management tools and technologies. This leads them to be able to apply these techniques to organize data, drive information from data, maintain high data quality, and practice Big Data Management on their own after finishing the course.

To reach the course objectives and ensure the proper learning, the course engages diverse activities including, watching videos, reading the materials, participating in quizzes, and completing the final project. This enables you to not only understand the basics of Big Data Management approaches but also apply them to develop a big data architecture, utilize big data management tools, and overcome real-world problems.

**Course Materials and Equipment (Prerequisite)**

To use this material you need to have a solid knowledge of Algorithm Design. Also, you should be familiar with Linear Algebra and Statistics. You are not expected to be an expert in coding. However, it is necessary to know some basics. Moreover, to complete the final project you need to install Apache Spark on your computer.

**Teaching and Learning Activities:**

1- Video Lectures: Presenting Concepts, Algorithms, and Mathematics as the basis of Big Data Management.

2- Reading Materials (Slides)

3- Virtual lab Project:

3.1- Problem representation: representation of a (real-world) problem.

3.2- Hands-on experience: executing Big Data Management methods on the problem using its tools.

4- Quizzes: after each chapter, a few questions (objective questions or short answer questions) are designed based on the content of the materials.

**Course activities:**

This course contains different types of activities which are listed below:

1) Reading the materials: for each chapter students are required to read the materials (slides) and get an intuition of the chapter, leading to understanding the content of the chapter.
2) Watching video lectures: students are expected to watch all the video lectures after reading the materials to better understand the content.
3) participating in quizzes: students should take part in quizzes and assess their understanding.
4) completing the final project: students should apply their acquired knowledge to manage a big data set using big data management tools and techniques.

**Course Objectives:**

The Big Data Management course will enable students to understand basic methods of big data processing, mining large and streaming datasets, extracting information, and using Big Data Management tools. During this course, students will learn how to use analytical tools and apply techniques to manage big datasets. Finally, students will apply their acquired knowledge and skills to build a Machine Learning project, manage a big dataset, and analyze it.

A summary of the course objectives are listed below:

1- Learning the basics of Big Data Management
2- Learning the theoretical ideas of Big Data Management, its concepts, and algorithms
3- Getting familiar with Data Mining and Data Stream Mining
4- Learning Big Data Management tools
5- Practice Big Data Management by using its tools on a Big Dataset

**Course Summary:**

There are nine main chapters of this course that guide students through modern techniques and approaches of Big Data Management, Data Mining, and executing Big Data projects on the Apache Spark platform.

This course broadens students knowledge and skills in the following areas:

1- Defenition and basics of Big Data Management

2- Data Mining and mining of Data Streams

3- Finding frequent items

4- Hadoop and Mapreduce

5- Spark and Big Data Management application

**Table of Contents:**

Week 1 Introduction to Big Data Models and Management:

Week 2 Introduction to Data Mining and Mining Data Streams, and a practical algorithm:

Week 3 Learning 2 important algorithms of Big Data:

Week 4 Learning one of the most important processing techniques of Big Data Management (MAp-reduce) and one of its tools (Hadoop) :

6.4- Combiners
6.5- Algorithms using Map Reduce

**7- Hadoop**
7.1- Introduction/ What is Hadoop and what is the use of it
7.2- HDFS
7.3- YARN
7.3- Quiz

### WEEK 5 LEARNING ONE THE MOST USED BIG DATA MANAGEMENT TOOLS IN DETAIL (APACHE SPARK):

**8- Apache Spark**
8.1- Sparks basics
8.1.1- Introduction to Spark
8.1.2- Spark Core
8.1.3- Resilient distributed Dataset (RDD)
8.1.4- DataFrames and Datasets
8.2- Ecosystem
8.3- Differences between Spark and Hadoop

### WEEK 6 (IMPLEMENTING A MACHINE LEARNING PROJECT WITH BIG DATA MANAGEMENT TOOL) :

**9- Lab**
9.1- Machine Learning with Spark
8.2.1- Introduction to MLlib
8.2.2- Hands-on project

**Laboratory Description and Equipment:**

The purpose of the virtual lab (final project) is to gain practical experience executing big data management techniques on a big dataset, implementing its algorithms, and working with its tools. Therefore, students should use Apache Spark Environment and complete the task.

**Course References:**

Tom White. Hadoop: The Definitive Guide. 4th Edition – O’Reilly. 2019.

Josh Wills, Sandy Ryza, Sean Owen, and UriLaserson. Advanced Analytics with Spark. O’Reilly Media. 2017.

Jure Leskovec, Anand Rajaraman, Jeff Ullman. Mining of Massive Datasets. 2011.

**Evaluation and Assessment Methods:**

Assessments take a variety of forms including Quizzes after each module (objective questions) or Implementation of a Big Data technique. To finish the course successfully, students are required to take 80% of the quizzes and complete the final project.

**The tasks lead to the production of the intellectual output and the applied methodology.**

There are four main steps, leading to the production of the intellectual output and the applied methodology:

1-Tasks

The Big Data Management course tasks are listed below:

1- Studying all the course references by the organization in charge (PTX).

2- Extracting information from the references based on the arranged content of the course.

3- Organizing and preparing slides, reading materials, and quizzes.

4- Developing video lectures.

5- Designing virtual laboratory (choosing appropriate Dataset and conducting codes)

2-Review and Feedback

All other partners will be asked to review the course materials and give feedback. This leads to enhancing the functionality of the course, improving its outcomes.

3-Fine-tunning

The responsible partner (PTX) should take the feedback into account and ameliorate the course materials.

4-Publishing

After the approval of the course by all partners, it will be published and is available for everyone online.

# 3 Course Contents