Presentation 3 / 20

COURSE OUTLINE

MODULE 12

Introduction to Big Data Hadoop and Spark
Introduction to Python for Apache Spark
Functions, OOPs, Modules in Python
Deep Dive into Apache Spark Framework
Playing with Spark RDDs
Data Frames and Spark SQL
Machine Learning using Spark MLlib
Deep Dive into Spark MLlib
Understanding Apache Kafka and Apache Flume
Apache Spark Streaming – Processing Multiple Batches
Apache Spark Streaming - Data Sources
Implementing an End-to-End Project

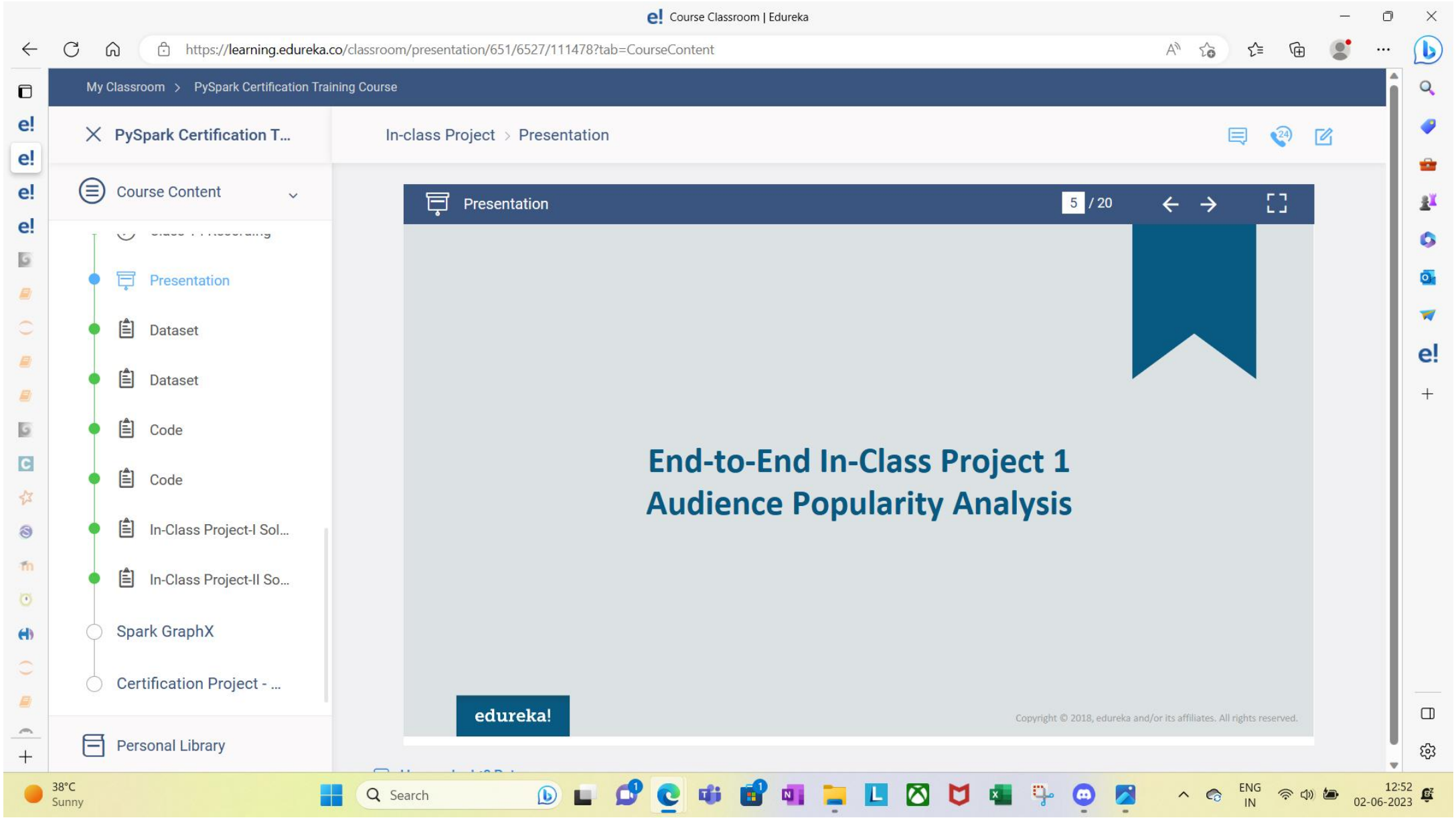
Presentation 4 / 20

Objectives

After completing this module, you should be able to:

- Create Spark based end-to-end Solutions for Domain Specific data using core Spark functionalities

edureka!


e! Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6527/111478?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
In-class Project > Presentation
Course Content
Presentation
Dataset
Dataset
Code
Code
In-Class Project-I Sol...
In-Class Project-II So...
Spark GraphX
Certification Project - ...
Personal Library
Presentation
5 / 20
End-to-End In-Class Project 1
Audience Popularity Analysis
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:52
02-06-2023

Presentation 6 / 20

Consideration

- You are going to appear for an Interview at World's largest Film and Media Production firm

edureka!

Presentation 7 / 20

Background Information

- You are going to appear for an Interview at India's largest Film and Media Production firm
- Since you have appeared as a candidate for Data Analyst profile you have to provide the firm with insights on the data that they provide

edureka!

Presentation 8 / 20

Dataset Overview

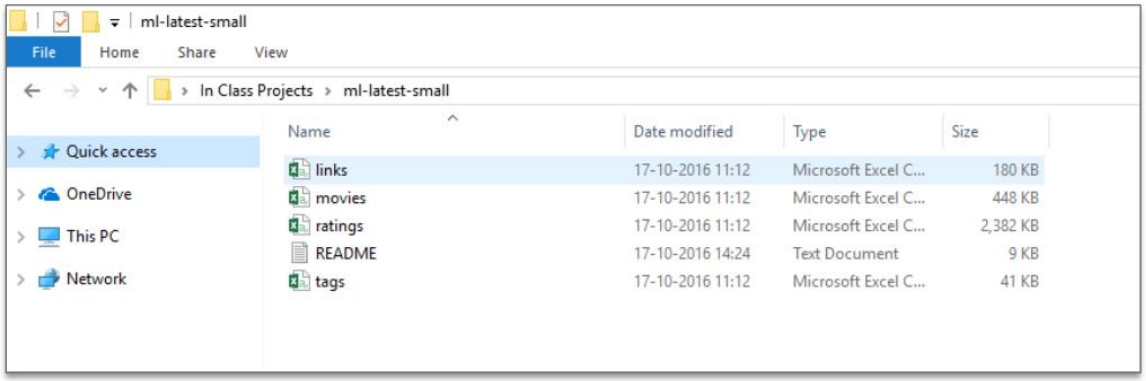
- You are going to appear for an Interview at India's largest Film and Media Production firm
- Since you have appeared as a candidate for Data Analyst profile you have to provide the firm with insights on the data that they provide
- In the next slide we will discuss about the dataset and the insights that you have to provide

edureka!

Presentation 9 / 20

Dataset

The given dataset has various files that you can download from here. You will only need the ml-latest-small.zip and ml-100k.zip files. Once you have downloaded the files your **ml-latest-small** download will contain some files like this:

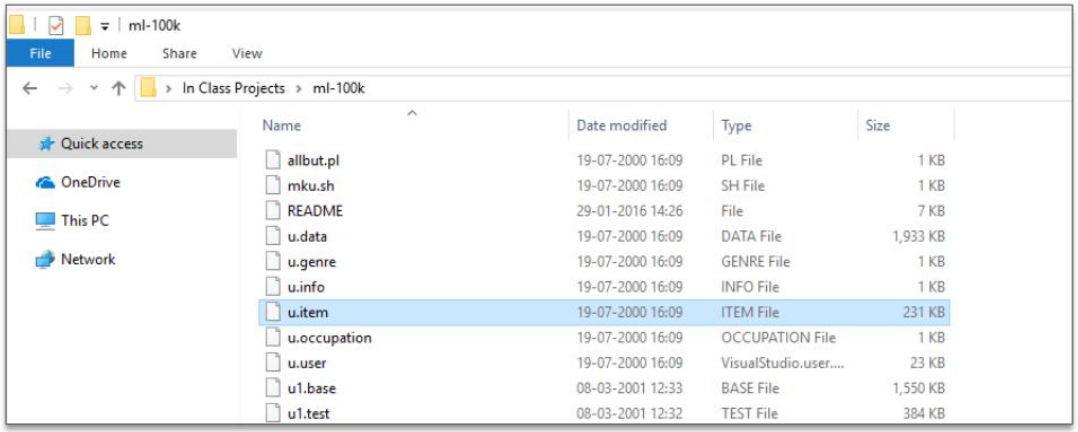


edureka!

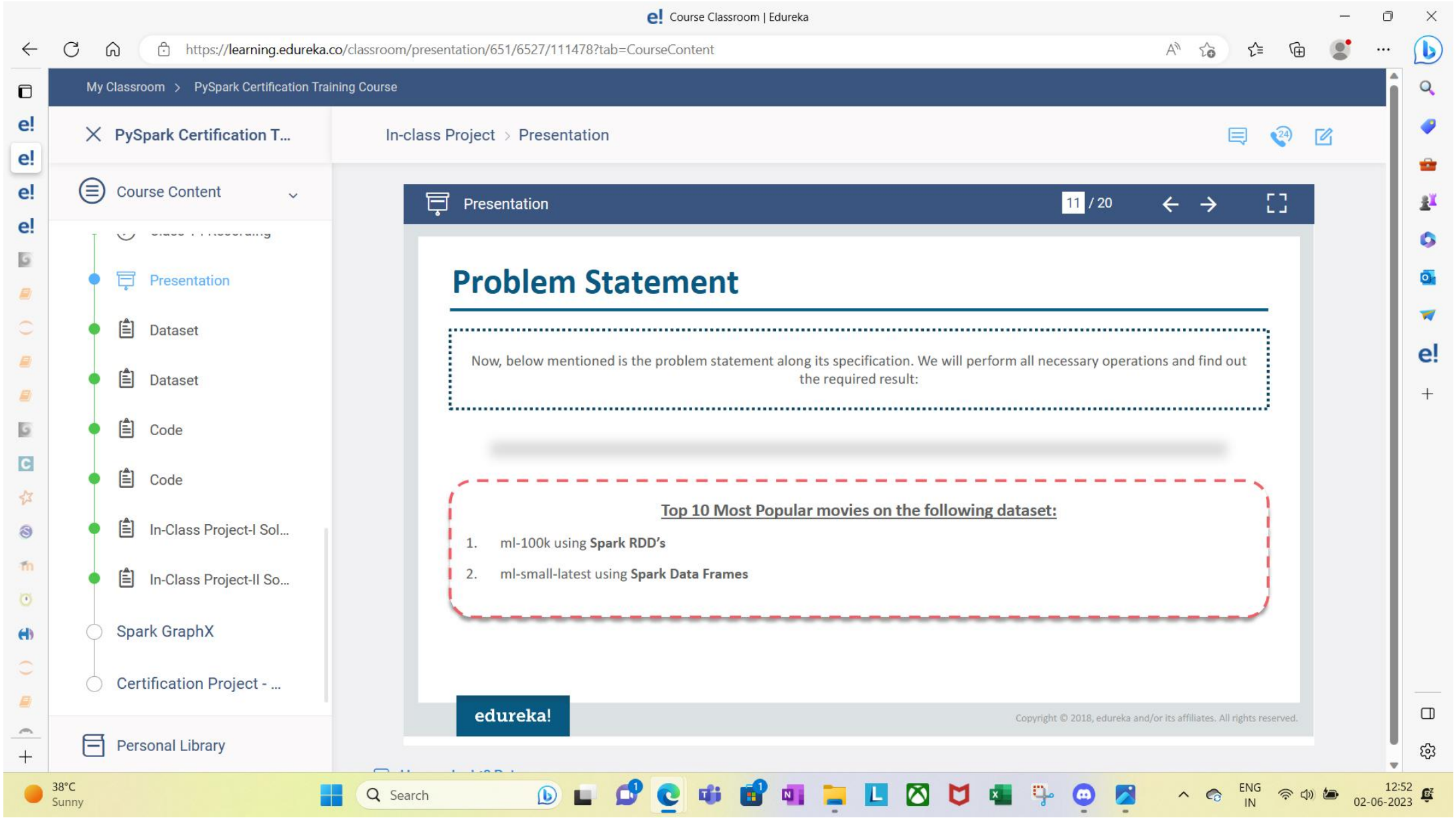
Presentation 10 / 20

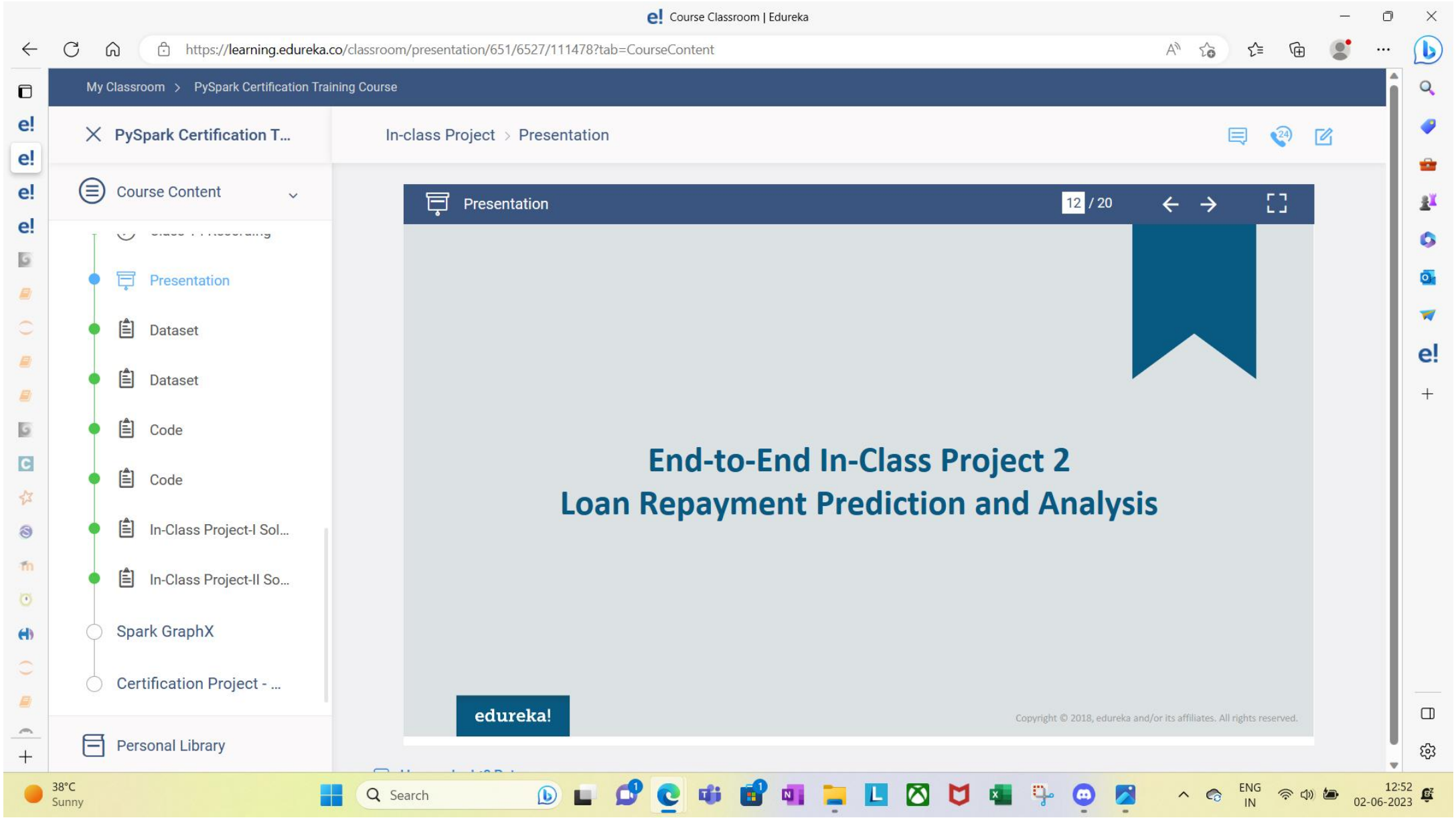
Dataset

This is how your **ml-100k** folder will look after it is downloaded. You may notice that previous folder contained CSV files while this has files with .data, item etc extensions. These files are pipe separated files and contain movie records of older releases.



edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6527/111478?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
In-class Project > Presentation
Course Content
Presentation
Dataset
Dataset
Code
Code
In-Class Project-I Sol...
In-Class Project-II So...
Spark GraphX
Certification Project - ...
Personal Library
Presentation
11 / 20
Problem Statement
Now, below mentioned is the problem statement along its specification. We will perform all necessary operations and find out the required result:
Top 10 Most Popular movies on the following dataset:
1. ml-100k using Spark RDD's
2. ml-small-latest using Spark Data Frames
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:52
02-06-2023


e! Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6527/111478?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
In-class Project > Presentation
Course Content
Presentation
Dataset
Dataset
Code
Code
In-Class Project-I Sol...
In-Class Project-II So...
Spark GraphX
Certification Project - ...
Personal Library
Presentation
12 / 20
End-to-End In-Class Project 2
Loan Repayment Prediction and Analysis
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:52
02-06-2023

Presentation 13 / 20

Consideration

- A leading financial bank is trying to broaden the financial inclusion for the unbanked population by providing a positive and safe borrowing experience

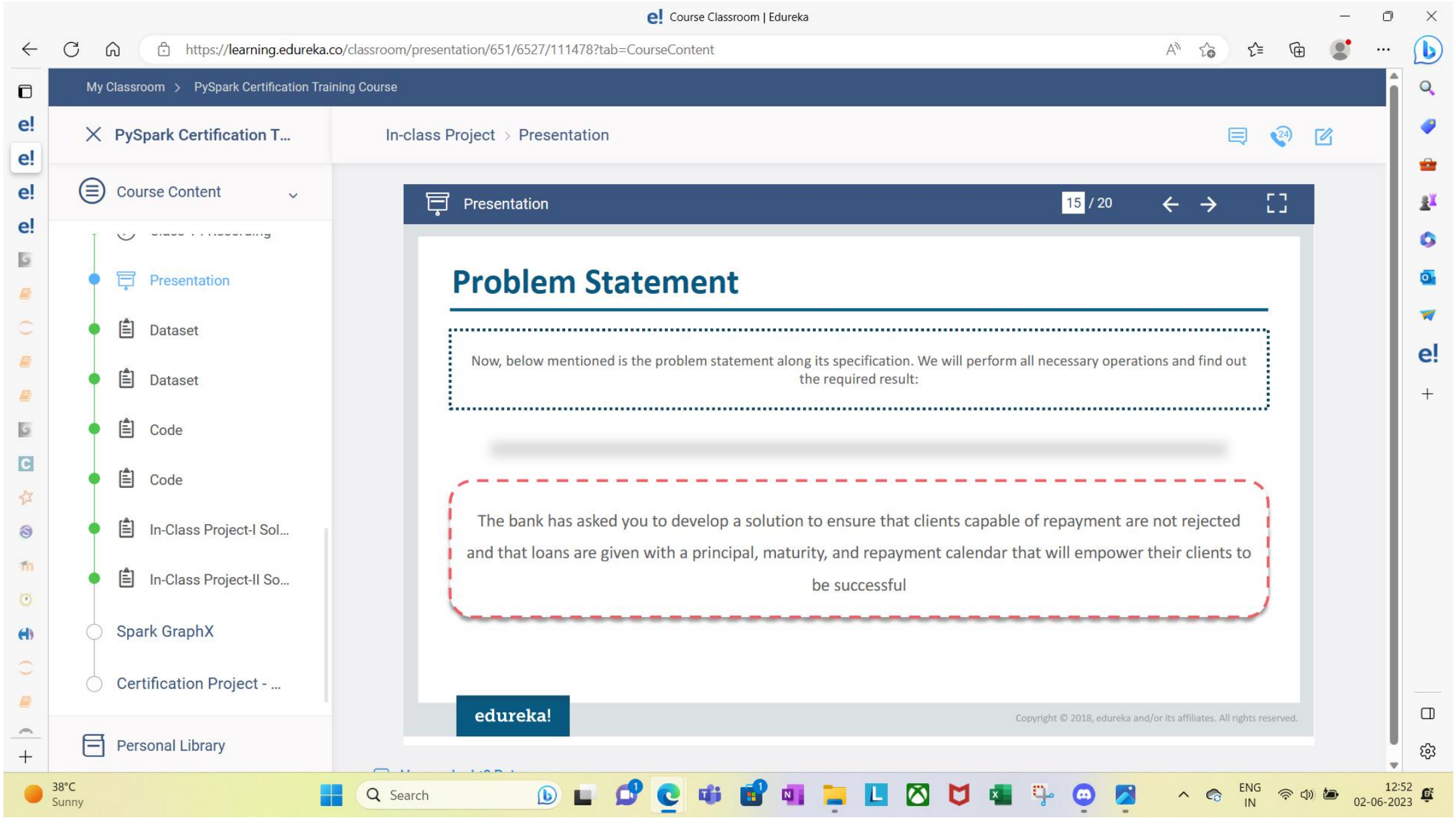
edureka!

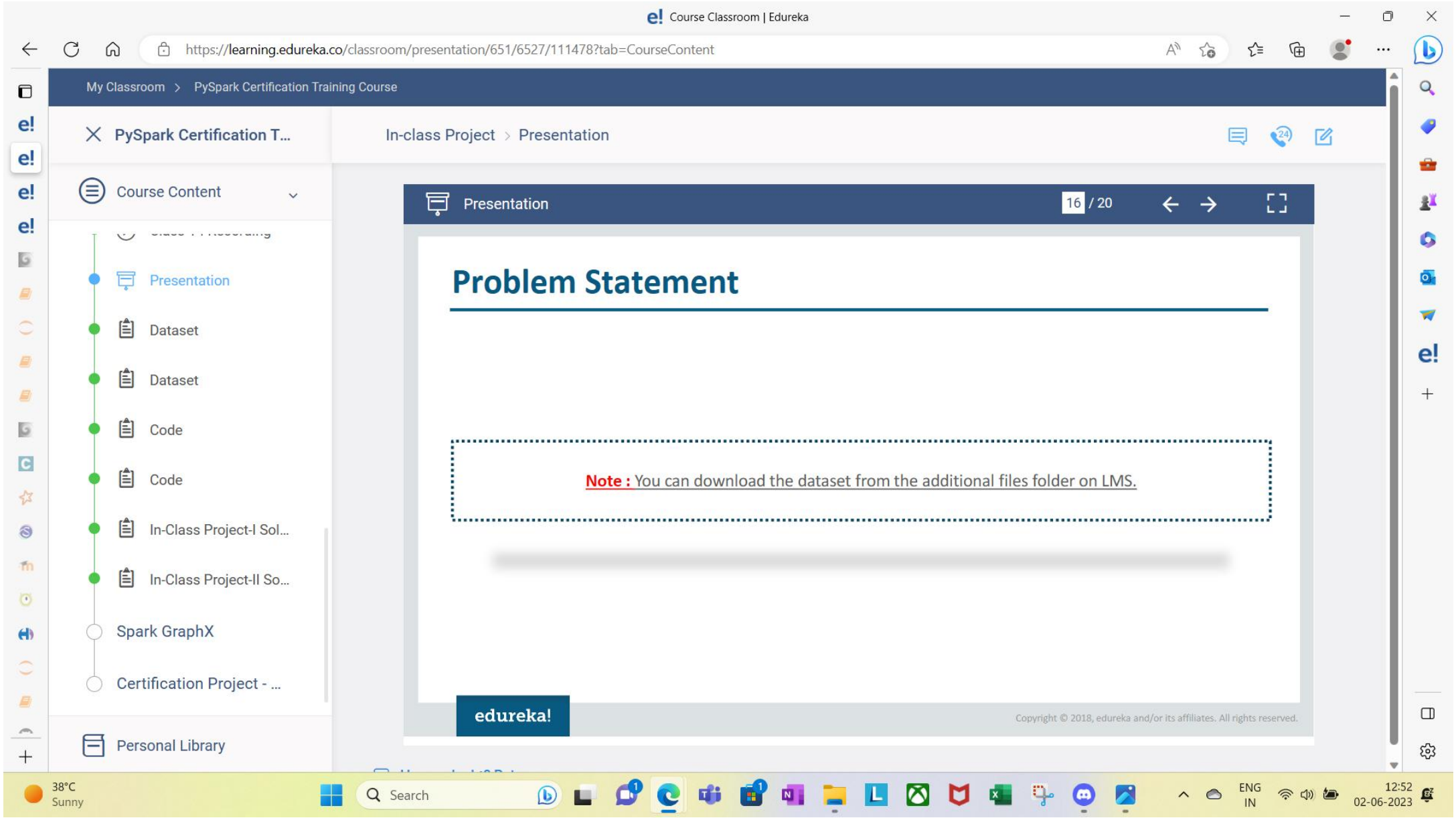
Presentation 14 / 20

Background Information

- A leading financial bank is trying to broaden the financial inclusion for the unbanked population by providing a positive and safe borrowing experience
- In order to make sure this underserved population has a positive loan experience, it makes use of a variety of alternative data--including telco and transactional information--to predict their clients' repayment abilities

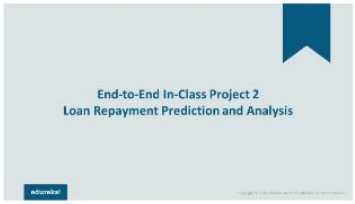
edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6527/111478?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
In-class Project > Presentation
Course Content
Presentation
Dataset
Dataset
Code
Code
In-Class Project-I Sol...
In-Class Project-II So...
Spark GraphX
Certification Project - ...
Personal Library
Presentation
15 / 20
Problem Statement
Now, below mentioned is the problem statement along its specification. We will perform all necessary operations and find out the required result:
The bank has asked you to develop a solution to ensure that clients capable of repayment are not rejected and that loans are given with a principal, maturity, and repayment calendar that will empower their clients to be successful
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:52
02-06-2023


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6527/111478?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
In-class Project > Presentation
Course Content
Presentation
Dataset
Dataset
Code
Code
In-Class Project-I Sol...
In-Class Project-II So...
Spark GraphX
Certification Project - ...
Personal Library
Presentation
16 / 20
Problem Statement
Note : You can download the dataset from the additional files folder on LMS.
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:52
02-06-2023

Presentation 17 / 20

Summary



edureka!