Project of TDT4305

BY HASSAN ABEDI

10.02.2022

Keywords: **Apache Spark**, **Big Data**, **Python**, **Databricks**, **Project**

Overview

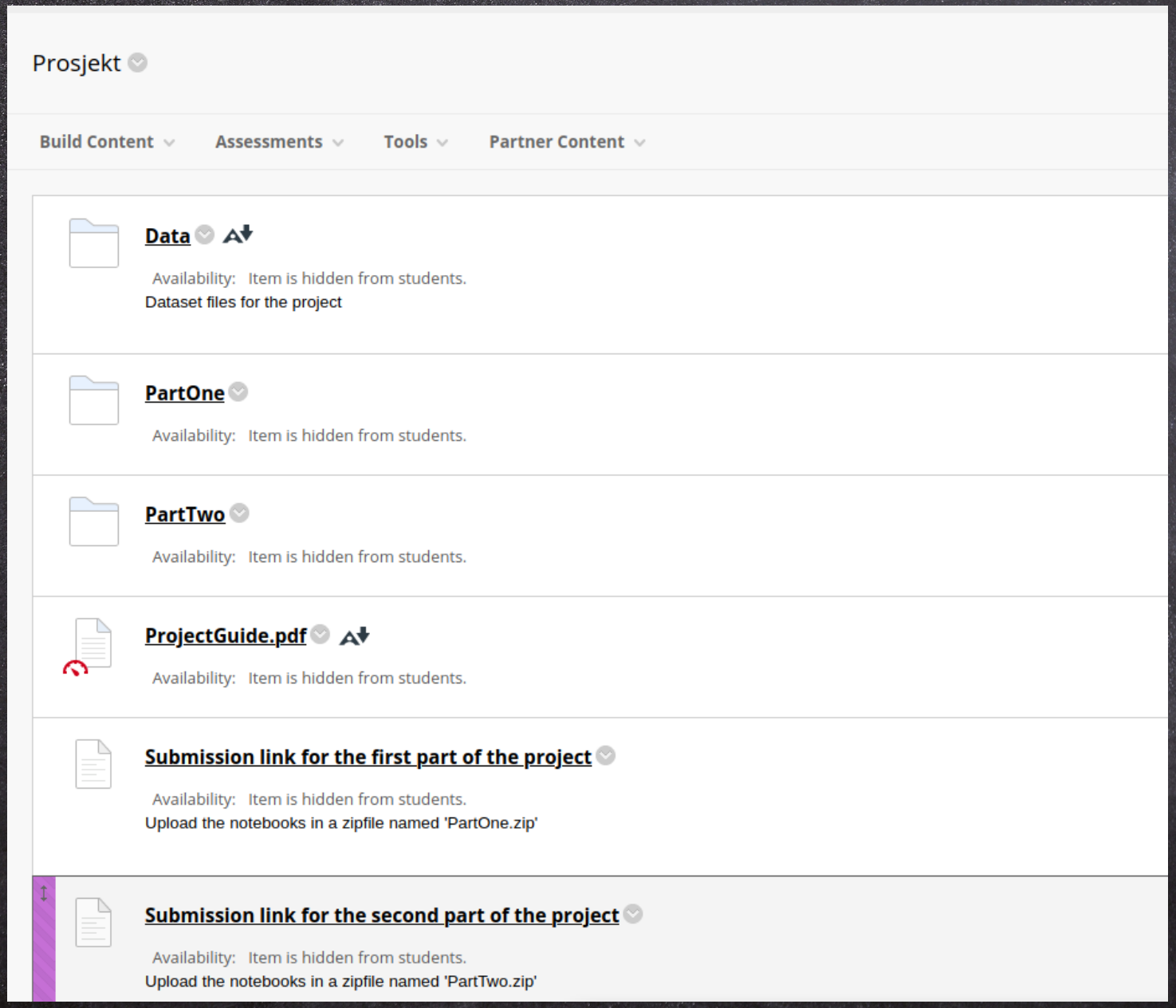
1. Assumptions
2. Project-related files
3. Structure of the project: parts and tasks
4. Tools and environment
5. Submitting your work
6. Communication method
7. A short demo of the environment

Assumptions

- Internet connection and a modern webrowser
- Have a Databricks Community Edition account (free to create)
- Know how to program in Python3 (intermediate level)
- Know how to write simple SELECT SQL queries
- Familiarity with Java, Markdown, and Jupyter is helpful but necessary

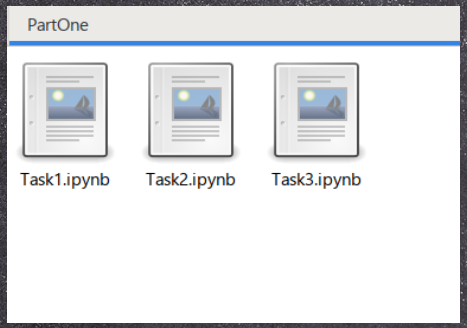
Project-related files

- All files are under 'Prosjekt' section on Blackboard
- Will become accessible today

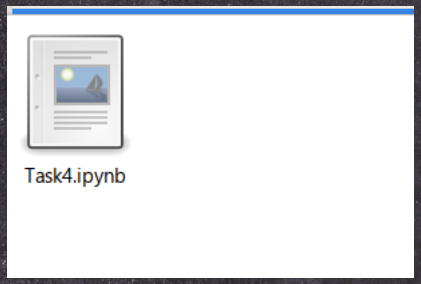


Structure of the project: parts and tasks

- Two parts
  - Part one
    - ⟶ Three tasks



  - Part two
    - ⟶ Only one task



- A task is a Jupyter notebook

Tools and environement

- We will use Databricks Community Edition
- To get started
  1. Create an account
  2. Sign in
  3. Create a Spark cluster
  4. Install the GraphFrames library on the cluster
  5. Import the data
- You can read the tutorial here
- Import the tasks' notebooks
- Read the content of each notebook carefully
- Limitations of Databricks Community Edition
  - After two hours of inactivity, a cluster will shut down and you'll need to delete it and create a new cluster; don't forget to reinstall libraries; the data won't be affected
  - Very rarely, you'll be logged off; so you'll need to sign in again

Submitting your work

- When done with a part
  - Download and save the notebook(s) on your machine
- Put the three notebooks for part one into a zipfile named 'PartOne.zip' and upload it via the submission link
- Put the notebook for part two into a zipfile named 'PartTwo.zip' and submit it
- **Weldone!**
- The deadlines for submissions and the grades for each part

Part	Submission Deadline	Grade Points
1	March 6, 2022, before 11.30pm	14
2	March 27, 2022, before 11.30pm	11

- Additional information
  - The most recent submission will be graded
  - Submission links are in 'Prosjekt' section on Blackboard
  - If working as a team, both people should submit separately

Communication method

- Please use Piazza for anything related to the project
- **Lykke Til!**