# Midterm Exam

# Big Data

Lecturer: Abdul Munif

**Friday, April 14, 2023, 08.00-10.00 (120 minutes)**.

**Method: OPEN BOOK (students can access online material, notes, or previous projects)**

## GENERAL RULES

- **Give the explanation and description for every step in your notebook.**
- **The marking will be based on the latest commit on GitHub. If you modify your repo and commit it later than April 14th, 2023, 10.am it will be considered as late submission.**
- **No discussion with other students in any possible way (chat, email, etc.)**

## PROBLEM

1. (Spark Streaming)
   Create an Apache Spark notebook for handling file stream inside a folder.
   - Use the data given in news folder (**news-1.json, news-2.json, news-3.json**)
   - Put the input inside folder named **input-your-student-id**. Example: input-51231132
   - Put the output inside folder named **output-your-student-id**
2. (Recommendation Systems – Frequent Pattern Mining)
   You are given the market basket dataset (**market-basket.csv**).
   - Use the FP-growth algorithm in Apache Spark to find the most frequent items. Save your result into an .xlsx files.
   - Do the experiment with different *minSupport* and *minConfidence* values. Give the conclusion at the end of experiment.

## SUBMISSION

- Submit all your notebooks into GitHub.
- Enter your GitHub link in the midterm submission.