Assignment 3 (20%)
STQD6324 Data Management
SEMESTER 2 2024/2025

Using the `u.user` file from the MovieLens 100k Dataset, which can be downloaded from https://grouplens.org/datasets/movielens/, write a Python script that functions as a wrapper to execute Cassandra Query Language (CQL) and Spark2 Structured Query Language (SQL) in order to answer the following questions. For each question, display only the top ten results:

i) Calculate the average rating for each movie.
ii) Identify the top ten movies with the highest average ratings.
iii) Find the users who have rated at least 50 movies and identify their favourite movie genres.
iv) Find all the users who are less than 20 years old.
v) Find all the users whose occupation is "scientist" and whose age is between 30 and 40 years old.

Your python script should include the following elements:

1. Python libraries used to execute Spark2 and Cassandra sessions.
2. Functions to parse the `u.user` file into HDFS.
3. Functions to load, read, and create Resilient Distributed Dataset (RDD) objects.
4. Functions to convert the RDD objects into DataFrames.
5. Functions to write the DataFrame into the Keyspace database created in Cassandra.
6. Functions to read the table back from Cassandra into a new DataFrame.

Optional: You may also attempt the above questions using HBase and MongoDB.

The deadline for submitting your script is **2025-06-19**. Please share your Jupyter Notebook with markdown via **GitHub**.

Criteria	**Marks**		
Reproducibility	3 The notebook is 100% reproducible	2 The notebook is reproducible with a few missing steps	**1** The notebook is not reproducible
Interpretation	**15** The interpretation of the findings is clear, easily understandable, and logical	**10** The interpretation of the findings is mostly clear and understandable, with minor areas needing clarification	**5** The interpretation of the findings is unclear and difficult to understand, lacking logical coherence
Overall GitHub presentation	2 The overall GitHub is i. properly structured, ii.each section neatly organized, iii. easy to follow	1 Part of the GitHub is i. properly structured, ii.neatly organized, iii. easy to follow	0 The GitHub is i. poorly structured, ii. each section is not organized, iii. hard to follow