

Assignment 4 (25%)
STQD6324 Data Management
SEMESTER 2 2023/2024

Using the *u.user* file from the MovieLens 100k Dataset (ml-100k.zip), which can be downloaded from <https://grouplens.org/datasets/movielens/>, write a python script that acts as a wrapper function to execute Cassandra Query Language (CQL) and Spark2 Structured Query Language (SQL) in order to answer the following questions [display only the top ten results for each question]:

- 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 with age that is less than 20 years old.
- v) Find all the users who have the occupation “scientist” and their 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 DataFrame.
- 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.

The deadline for submitting the script is **11 June, 2024**.

Please submit to bernardlkb@ukm.edu.my.