CPC451/CPM451 Big Data Technologies and Management
Semester 2, 2023/2024

PROJECT (20%) – REPORT & PRESENTATION
(Group Work: Up to 4 members per group)

Instruction: The project will be evaluated based on group work and as well as individual performance via a written report and group presentation. Every group must submit a written report and provide a group presentation. Group formation is conducted via the elearn portal.

Deadline: Tuesday, 25th June 2024 (5:00 pm), submit your softcopy of your report/slides through CPC451/CPM451 e-learning portal during class time. Group presentation will be conducted in the class for two weeks. Random drawing will be made to determine which group to present in the first week (25/06/2024) & (28/06/2024) and second week (2/07/2024) & (5/07/2024).

Project Description: Each group should identify a big data set which is between 400MB to 500MB to be used in Stand-alone MongoDB implementation and cloud based (such as MongoDB Web or Google Cloud) / Container (such as Docker). The datasets can be obtained from https://www.kaggle.com/datasets and other open dataset repositories. Build MongoDB a database and populate the chosen dataset in MongoDB both stand-alone and MongoDB Web/Container. You may need to download and install MongoDB and use the Mongo Web cloud services/Container. Enter the data set into the databases. Run at least **four** meaningful queries that are best describing the data inside the database. Meaningful queries **exclude** CRUD (Create, Read, Update, Delete) operations. Compare and discuss their performance in terms of ease of use, creating queries and data processing speed.

Your deliverables must include the following requirements:

- Chosen platform for MongoDB database: stand-alone and cloud/container
- Installation process and data entry
- At least four meaningful queries in MongoDB
- The queries have to be identical (same) for both in stand-alone and cloud/container
- Compare and discuss their performance
- Observation & lesson learned

2. Below are some points that guide you in preparing the report.
   - *i. Abstract*
   - *ii. Introduction*
   - *iii. Project Content*
     1. *Brief description of the given dataset (about, size, why it is chosen)*
     2. *Selection of implementation platform for MongoDB*
     3. *Installation process, database construction and data entry*
     4. *At least 4 meaningful queries in MongoDB (stand-alone)*
     5. *At least 4 meaningful queries in MongoDB (cloud/container)*
     6. *Comparison, discussion and observation*
     7. *Concluding remarks*
   - *iv. Lesson learned from the project*
   - *v. Clear division of group members' roles*

Deadline for submission is 25th June 2024, by 5:00 pm. Online submission via elearn.

vi. *Conclusion*
vii. *References (At least 4 references)*
viii. *Appendices (If any)*

3. Marking Scheme (provided):
    - Marking Scheme: refer to the rubrics posted on the e-learning page

4. For the in class presentation/demo, each group is allocated about 15 minutes including Q & A:
    - Everyone in the group is expected to present some portion of the project
    - Only demo one meaningful query for each in stand-alone and cloud/container
    - Comparison all four queries in both types of implementations

5. Submit the following together with IEEE formatted report (**One** submission per group):
    - Soft copy - (Report and slides): e-learning
    - IEEE format (refer to the elearn for the sample template)

Note:
The report should include an appendix indicating detailed descriptions on contributions of each group member in the project. *In the event that parts of the report are directly copied from others without references, F grade is given.*

Deadline for submission is 25th June 2024, by 5:00 pm. Online submission via elearn.