Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA
Cawangan Perlis, Kampus Arau
Group Project
DSC650 - Data Technology and Future Emergence
Mar 2025 – Jul 2025 (20252)

LEARNING OUTCOME:

- Analyse existing big data sets using data technology tools (C4)

SUBMISSION DEADLINE: Week 14 (as in the UF assignment due date)

Each group should consist of FOUR (4) members. Each group needs to develop a big data project utilizing Hive or Spark.

Write project report based on the following report outline:

1. Introduction
   1.1. Background
      - Provide an overview of the selected problem or scenario.
   1.2. Objective
      - Example: Analyse the decline in product sales performance.
2. Requirements Analysis
   2.1. Source the dataset
      - Source of data dataset can be obtained from open data platforms such as Kaggle, Data.gov, or Portal Data Terbuka.
   2.2. Data analysis requirements.

      Define **at least 10 queries or predictions** for each project.
      Examples:
      - Predict top-selling products for the next quarter.
      - Analyze customer churn based on purchase patterns.
      - Determine the busiest hours of sales across cities. \
      - Predict sales trends using machine learning models in PySpark.
      - Detect anomalies in sales data.
   2.3. System Requirements
      - Specify the big data tools/platforms used (e.g., Hive for querying, Kafka for real-time streaming, Spark for analysis).
      - Mention hardware/software requirements.
3. Database Design
   3.1. Data Model
      - Develop an **ERD** for structured data or a **NoSQL schema** for unstructured data.
   3.2. ETL Process
      - Describe the Extract, Transform, and Load (ETL) process.
   3.3. Real-Time Data Pipeline (Optional)
      - Implement a streaming data pipeline using Kafka or Spark Streaming if applicable.

4. Implementation
   4.1. Data Preprocessing and Transformation
      - Detail how the dataset was cleaned and transformed for analysis.

4.2. Data Analytics and Machine Learning
   - Implement data analysis and prediction queries using tools like HiveQL or PySpark.
   - Include visualization of key insights using Python (Matplotlib/Seaborn) or Power BI.

5. Conclusion
   - Does this project successfully achieve its goals? Describe the project's outcome.
   - lesson learned from the database project implementation.
6. References
   - Cite all datasets, tools, and libraries used in the project.