Graduate Certificate in Artificial Intelligence with Machine Learning
AIGC 5503 – AI For Bus. Decision Making

Lab 8 & 9: Case Study/Hands-on with Apache Spark

Submission guidelines:

- For this lab, you will need to submit your answer through Blackboard.
- Submit your response to the Module Activity in the Blackboard discussion board
- Submit 1 PDF file for your implemented code.
- Name the PDF as follows: firstname_lastname_LAB9.pdf
- Go to the course Blackboard → Labs folder → Lab Exercise 8 & 9 and submit the pdf.

Lab goals:

- Analyze real-world case study for the application of Apache Spark & Spark Streaming.
- Implement and Analyze using Apache Spark.

Part 1: Module Activity Participation (Module 10 + 11)

- Complete the activity **Module 10 – Self-Research**
-

Reflect on the topics we've covered and choose one of the following questions to discuss on the discussion board:

Questions (Choose One):

1. **Spark's Unified Computing Engine:**
    - Reflect on the value of Spark's unified engine (batch, streaming, SQL, ML) in business AI systems. How does this architectural design improve efficiency and agility in enterprise decision-making workflows?
2. **Structured APIs and Business Data Analytics:**
    - Consider how Spark SQL, DataFrames, and Datasets facilitate high-performance, scalable analytics. In what types of business scenarios would these structured APIs be most impactful? Provide an example or hypothetical use case.
3. **MLlib and Predictive Modeling in Business Contexts:**
    - Discuss the benefits and limitations of using Spark MLlib for predictive analytics in business (e.g., churn prediction, fraud detection, inventory forecasting). What considerations should a data team have when selecting Spark MLlib over other ML libraries?

Write a thoughtful response to the question you select.
Post your response on the discussion board.
Engage with at least one peer's post by providing additional insights or asking a follow-up question.

- Complete the activity **Module 11 – Self-Research**

Spark streaming is one solution for steaming that we have discussed. In this activity, you will research alternative solutions to streaming big data for Real-Time Business Intelligence
Steps:

1. Find and select a Streaming Tool to process big data (i.e. Kafka, Flink, NiFi, etc).
2. Summarize the details of the Streaming Tool:
    a. Overview
    b. Comparison between Spark Streaming
    c. An example of a Real-World Use Case

Post your findings on the discussion board
Engage with at least one peer's post by providing additional insights or asking a follow-up question.

Part 2: PySpark Implementation

- Review the online tutorial on setting up PySpark for Machine Learning.
    - **Spark Tutorial**

- Reproduce the analysis conducted in previous labs on the OnlineRetail.csv dataset.
    - Perform the following analysis
        - Data Exploration Analysis
        - RFM Analysis
        - Clustering Analysis

- Submit a copy of your code with the generated outputs & plots in the form of a PDF file.

Evaluation :

- **Lab 8**
    - Part 1 (Module 10 Activity) = 50%
    - Part 2 (Module 11 Activity) = 50%
- **Lab 9**
    - Part 2 – 100%

Enjoy!