**Task 1:**

*Focus on manipulating data using Spark RDD, Spark DataFrames and Spark SQL*

* **Part 1:** Work with RDDs in PySpark to implement specific queries related to flight delays data analysis.
* **Part 2:** Work with Dataframes in PySpark to implement specific queries related to flight delays data analysis.
* **Part 3:** Implement a query and compare three different approaches: RDD, DataFrame, and SparkSQL.

**Task 2 - Part A: Preprocessing, Visualization, Data Modelling**

*Use Spark DataFrames, Spark SQL, MLlib/ML packages to do the pre-processing, visualization, and data modelling steps. The data modelling task aims to create the machine learning models utilizing MLlib/ML APIs to predict both departure and arrival delays from the testing data using classification tasks. Build models to predict the classes.*

* **Part 1:** Data Loading, Cleaning, Labelling, and Exploration
* **Part 2:** Feature extraction and ML Training

**Task 3 - Part A: Real-time stream processing on big data**

*Simulate the streaming data production using Kafka, then show some basic streaming classification to display the prediction accuracy and total number of flight records after consuming the data. Build a streaming application that integrates the machine learning model that can classify the flight-delays data stream.*

* **Task 1:** Kafka Producer - Producing the streaming data, where you can use csv modules to read and publish the data to the Kafka stream.
* **Task 2:** Kafka Consumer - Consuming the streaming data using Kafka consumer, process and visualize the ingested data from Kafka.
* **Task 3:** Streaming application - Using Spark Structured Streaming together with Spark ML/SQL to process data streams.