# Assignment 4
# Spark MapReduce and DataFrames

In this programming assignment, we will practice using Spark to perform both MapReduce with RDDs and tabular data analysis with DataFrames. You will be working with two data files;`flights_data.txt` and `Combined_Flights_2021.csv`

## [20 Points] P1: MapReduce with Spark RDDs

You are given the file **flights_data.txt** containing information about flights, including the flight date, airline name and origin airport for each flight. Each line in the dataset represents a flight record, with fields separated by commas. Your task is to develop a MapReduce program to identify **pairs of airlines that operate on the same date.** In other words, you need to find the **pairs of airlines that share the same origin and the same dates**, and determine the count of each pair. You must sort the airline **pairs alphabetically (a-z)** (for both airline names in the pair) with the **counts in descending** order.

**** Don't filter out canceled flights**

**** Pairs (A, B) and (B, A) are equivalent, only count once.**

**** If (A,5) and (B,3) then ((A,B),3)**

**** Don't report self-pair = eg., (A,A) is self pair**

Implement the function `airline_pairs_by_origin()` in the given template. Your implementation should be in Spark using RDDs and the MapReduce paradigm. You might need to use multiple Map and Reduce operations. For (each) `map()`, determine the type of the output **key-value pairs**. For (each) `reduce()/reduceByKey()`, determine the output type. For example, for the **word count (2)** exercise in Lab 8, the output types are `[(str, 1)]` and `[(str, int)]`.

**Add the output type as a comment line in your code script before each operation such as (map, reduce, filter, …) .**

**Note: For this part you should not use Pandas DataFrame.**

Example Input:

```
flight_date, airline, origin
2021-01-01, Airline_A, Org_1
2021-01-01, Airline_A, Org_1
2021-01-01, Airline_A, Org_1
2021-01-01, Airline_B, Org_1
2021-01-01, Airline_B, Org_1
2021-01-01, Airline_C, Org_1
2021-01-01, Airline_A, Org_2
2021-01-01, Airline_B, Org_2
2021-01-01, Airline_C, Org_2
```

Example Output:

```
[((Airline_A,Airline_B),3),
((Airline_A,Airline_C),2),
((Airline_B,Airline_C),2)]
```

# [80 Points] P2: Data Analysis with Spark DataFrames

Given the file `Combined_Flights_2021.csv` You will implement the four queries given using **Spark DataFrames**. Specifically, implement the following:

- [20 Points] What is the **name** of the airline that had the most canceled flights in January 2021? Implement the method `air_flights_most_canceled_flights()`.
  **`Return only the name as string.`**

- [20 Points] How many flights were diverted between the period of 1st-30th November 2021? Implement the method `air_flights_diverted_flights()`.
  **`Return only the number of flights as an integer number.`**

- [20 Points] What is the average airtime for the flights from “Los Angeles, CA” to “New York, NY”? Implement the method `air_flights_avg_airtime()`.
  **`Return only the average airtime as a float number.`**

- [20 Points] How many unique days are missing departure time (DepTime)? Implement the method `air_flights_missing_departure_time()`.
  **`Return only # unique days as an integer number.`**

The `Combined_Flights_2021` CSV file can be found on Kaggle.

For both problems, you are given the code template `assignment4_template.py`. Your task is to fill in the missing code indicated by a raised `NotImplementedError`.



# Submission Instructions

- The assignment is due at 11:59PM on Sunday, November 23, 2025.
- Your code must be in Python within the provided template. **Any modifications to the template (method signatures, main function, etc) will incur a 10% penalty.**
- Your submission should be a single python script of the filled-in template with the following name format: `<first_name>_<last_name>_<ID>_A4.py`

  **There should not be any space in filename, only underscores are allowed**

  (e.g. `john_doe_11111111_A4.py - Valid`

  `john doe_11111111_A4.py - Invalid`) .

  Do not zip the file or provide explanations in pdf/text files.
- If you need clarification about an unclear part of the assignment, send an email to your respective TA.
- If you require help in programming, please schedule a POD session with your respective tutor and prepare your questions. The tutors may assist you with the programming and APIs but will not provide solutions to the assignment.
- This is an **individual** assignment. You are not allowed to copy/share your solutions with your colleagues. Doing so is considered cheating that disqualifies both submissions (0%) and may be reported to the department.

# Late Policy

- 0-24 hours late = 25% penalty.
- 24-48 hours late = 50% penalty.
- More than 48 hours late = you lose all the points for this assignment.
- **Submissions of corrupted files, blank files, or the assignment template will be considered late submissions.**