Université Concordia University

Department of Computer Science & Software Engineering
COMP6231 Fall 2022

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.

[40 Points] P1: MapReduce with Spark RDDs

You are given the file `worker_shifts.txt` containing restaurant workers' shifts. Each line contains a worker's name and a shift (day) he/she worked in, separated by a comma. Your task is to find the co-workers who worked together on the same day most often. Implement a MapReduce program that finds the pairs of workers who worked together for at least one shift and the number of shifts they worked together.

Implement the function `restaurant_shift_coworkers()` 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 9, the output types are `[(str, 1)]` and `[(str, int)]`. Add the output type as a comment before each map or reduce call.

Example Input:

```
Fabian Henderson,2022-01-18
Fabian Henderson,2022-01-19
Fabian Henderson,2022-01-20
Fabian Henderson,2022-01-21
Shreya Chmela,2022-01-19
Shreya Chmela,2022-01-20
Shreya Chmela,2022-01-21
Shreya Chmela,2022-01-24
Leila Jager,2022-01-23
Leila Jager,2022-01-24
```

Example Output:

```
[(('Shreya Chmela', 'Fabian Henderson'), 3),
(('Fabian Henderson', 'Shreya Chmela'), 3),
(('Shreya Chmela', 'Leila Jager'), 1),
(('Leila Jager', 'Shreya Chmela'), 1)]
```

[60 Points] P2: Data Analysis with Spark DataFrames

Given the file `Combined_Flights_2021.csv` we studied in Assignment 2, you will implement the four queries given but using Spark DataFrames. Specifically, implement the following:

- [15 Points] What is the name of the airline that had the most canceled flights on September 2021?
  - Implement the method `air_flights_most_canceled_flights()`
- [15 Points] How many flights were diverted between the period of 20th-30th November 2021?
  - Implement the method `air_flights_diverted_flights()`
- [15 Points] What is the average airtime for the flights from “Nashville, TN” to “Chicago, IL”?
  - Implement the method `air_flights_avg_airtime()`
- [15 Points] How many unique days are missing departure time (DepTime)?
  - Implement the method `air_flights_missing_departure_time()`

The CSV file can be found on Kaggle. Alternatively, you can find the file here.

For both problems, you are given the code template `assignment4.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 Wednesday, December 07, 2022.
- 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` (e.g. `john_doe_11111111_A4.py`). 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 mossad.helali@mail.concordia.ca.
- 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.**