Assignment

adsb.json **oag.json**

Objective

We are excited to present you with a technical assignment to assess your skills in data science, specifically for working with large-scale data processing, your knowledge of technologies such as Apache Spark.

Overview

You are developing a system to process and analyze airport and flight data in real-time. Your objective is to design and implement a scalable solution for analyzing large datasets generated by various sources - e.g. ADSB and OAG.

Data Processing with Apache Spark

1. Load a dataset simulating airport and flight data - `adsb.json, oag.json` files.
2. Use Apache Spark to ingest and process the data (e.g., data cleaning, aggregation, transformation).
3. Conduct simple analysis. Compute some basic airport KPIs, including but not limited to:
   a. average speed for each airport
   b. the total number of delayed flights (categorized into arrival delays and departure delays)
4. Filter and transform a DataFrame by applying a window function (Spark partitioning):
   a. Filter the DataFrame to retain only the most recent entry (the one with the smallest `LastUpdate`) for each `FlightId`.
   b. Return a DataFrame containing only the `FlightId` and the corresponding latest `LastUpdate`.
   c. Example:

FlightId	LastUpdate
BL400	2023-10-03 06:00
BL400	2023-10-03 07:00
BA484	2023-10-03 08:00
BA484	2023-10-03 09:00

6. Present the results in the form of a graph or as a table.

Submission of Results

Requirements

1. **Code Submission:** Push your code to a (**GitHub repository** - public or private) and provide us with the repository link and necessary access permissions.
2. **Documentation:** Feel free to include `README.md` file in the repository with additional and useful information.

Guidelines

- **Clean Code:** Ensure your code is organized, with clear function definitions and error handling.
- **Formatting:** Code is consistently formatted and follows Python conventions.
- **Dependency Management:** Provide a `requirements.txt` for easy installation of dependencies.
- **Error Handling:** Make sure your script handles errors and missing data.

Please read the instructions carefully and ensure all requirements are met. If you have any questions, feel free to reach us out.