Assignment 2 (25%)
STQD6324 Data Management
SEMESTER 2 2023/2024

Airline on-time performance

Have you ever found yourself stranded in an airport due to a delayed or cancelled flight, pondering if the situation could have been anticipated? This project offers an opportunity to explore such scenarios.

In this project, you will use data from Kaggle: https://tinyurl.com/u8rzvdsx, encompassing airline performance data from 1995 to 2007. Each participant need to choose only one year for analysis.

The challenge:
Find answers to the following questions based on the data available on the website:

- What are the optimal times of day, days of the week, and times of the year for minimizing flight delays?
- What are the primary factors contributing to flight delays?
- What factors predominantly lead to flight cancellations?
- Which flight experiences the most frequent and significant delays and cancellations?

For each question, utilize **Pig** or **Hive** to extract insights from the dataset and generate figures to explain your findings using **Python or R**. Please use your creativity to answer these questions, and you are encouraged to search online for ideas of how others have tackled similar challenges, such as on this link: https://tinyurl.com/bdejna9e.

Please ensure that the scripts and codes used to generate your findings are included within the main report. You can use the Google Colab for this purpose. The submission deadline is **May 31, 2024**. Please share your work through GitHub.

1995.csv → A`ZRA ZULAIKHA
1996.csv → PAN LUOCHUAN
1997.csv → MOHAMMAD WAFIUDDIN
1998.csv → NAZMI AZIM
1999.csv → NUR DIANA
2000.csv → IRFFAN HAZIQ
2001.csv → ATHIRAH
2002.csv → LIU XIAOTIAN
2003.csv → PAN ZHANGYU
2004.csv → MOHAMAD RADZMI
2005.csv → SUN YUCHEN
2006.csv → LI YUTONG
2007.csv → KAMARUL ARIFIN