# INFO7250 Project Proposal

Student ID：001569847

Student Name：Jialiang Xiao

## Dataset

The dataset I chose for this project is Airline Delay Analysis. It is a Kaggle dataset, and its link is <https://www.kaggle.com/sherrytp/airline-delay-analysis>.

The datasets contain daily airline information covering flight information, carrier company, taxing-in, taxing-out time, and generalized delay reason of exactly ten years, from 2009 to 2019. The data for 2020 is not complete. The DOT's database is renewed from 2018, so there might be a minor change in the column names.

## Analysis Plan

### Analysis with MapReduce

I plan to use MapReduce to do the following analysis based on the dataset.

Count the number of delays of each op\_carrier(airway company) from 2009 to 2019.

Count the number of delays of each op\_carrier for each year.

Count the number of delays of each op\_carrier for each month from 2009 to 2019.

Find the date of the highest number of delays for each op\_carrier.

Find the hour of the highest number of delays for all op\_carrier from 2009 to 2019.

Find the hour of the highest number of delays for each op\_carrier from 2009 to 2019.