Data Engineering - Project Proposal

By Mugdha Dixit, Sushmitha Mudduluru, Yvette Iradukunda

1. Whether you are working individually or as a pair and the names of your team members (max 3 people with express consent from the professor).

   Team Members:

   - Mugdha Dixit
   - Sushmitha Mudduluru
   - Yvette Iradukunda

2. Provide links to the public data source(s) you plan to use for this project.

   Flight Status Prediction:
   https://www.kaggle.com/datasets/robikscube/flight-delay-dataset-20182022

3. Make the case that the dataset satisfies one or more of the differentiating V's of big data (volume, velocity, or variety). Again, volume and velocity do not need to be in the order of terabytes to be suitable here.

   Volume: 4 GB.
   Variety: It's a structured dataset

4. Provide a general problem that you will work on with respect to this dataset and the big data concepts and tools that you might use to solve it.

   - Using this "Flight status prediction" dataset we are planning to explore how different airlines compare.
   - We will be using Apache Spark for this project.

5. Provide three real-world questions that you would like to answer regarding the dataset that you chose.

   Following are some real-world questions we would like to solve:

- How many flights are delayed in that particular year and which airline are those.
- Which airlines have the highest rate of flight delays and cancellations in the US?
- What are the most common causes of flight delays in the US and how have they changed over time?