**INFO 7250 ENGINEERING OF BIG DATA SYSTEMS**

**PROJECT PROPOSAL**

**Submitted by: Venkata Nitin Reddy Byreddy**

The dataset that I am going to use is the data on **Airline On Time Data.**

Link to the dataset: <https://community.amstat.org/jointscsg-section/dataexpo/dataexpo2009>

This dataset contains of information about the airlines departure and arrival time.

I chose this dataset as it is huge and has many number of columns which will be helpful to perform various map reduce programs, pig, hive for analysis.

**I will try to analyze the following questions:**

1. When is the best time to fly to minimize delays?
2. What are the causes of delay?
3. How well does weather predict delays?
4. Which airline is the busiest?

I hereby declare that this project is my own work, that is authentic, and no part of this project has been taken any online or offline resources.

If this project is found not to be authentic, I take full responsibility.

And, I have read the guidelines as stated on the following page:

<https://osccr.sites.northeastern.edu/academic-integrity-policy/>

And, I am aware of the consequences of not complying with the policy.

Venkata Nitin Reddy Byreddy