INTRODUCTION
Flights have become commonplace in our lives but the efficiency is not as good as it can be. Gate conflicts, operational challenges, air traffic management, dynamics of a flight can change quickly and lead to costly delays. Airlines spend around 22 million dollars on improving the efficiency. Real time big data analysis is the solution to this problem. Pilots could augment their decision making process with this real time information given to them.

To reduce the delays and to make the decision making more accurate and efficient, we have come up with a module which will predict the delays in flight considering various factors. Large real time data related to flights, airports, weather are combined, analyzed and made into a model. Future predictions are made based on this model.

OBJECTIVES
Main Objective
• Help in pilots decision making by providing real time business intelligence by predicting delays.

Secondary Objective
• Remove noise from the data.
• Collect and combine relevant flight data to build a model.
• Extract the important parameters that contribute to delays.
• Train the model by classifying the data into different classes defined by their delay time.

RELATIONAL MODEL OF DATA
• Understanding and cleaning the data
• Building a relational structure into the data
• A Robust mapping algorithm to handle numerous weather nominal values
• Transforming data based on PCA
• SVM not settling down due to non-convex nature of input data

CHALLENGES
• Original data size : 211,532 samples and 35 features
• After cleaning data: 165,021 samples and 35 features
• After joining related data: 165,021 samples and 23 features
• After feature extraction using PCA : 165,021 samples and 13 features

RESULTS
Feature Extraction Results:
• Original data size: 211,532 samples and 35 features
• After cleaning data: 165,021 samples and 35 features
• After joining related data: 165,021 samples and 23 features
• After feature extraction using PCA: 165,021 samples and 13 features

Classification Results:
• Accuracy was only 28.30%.
• The classification on test data was very poor, because training data was locally grouped (example: data specific to one airport was stored together)

CONCLUSIONS
After collecting the flight related data over a long period of time and analyzing the data, we have constructed an approximate model to predict delays on particular route given the current air traffic, weather and other conditions.
This is a definite improvement over current systems and the real-time information produced is valuable in re-calculating routes for critical flights and to inform the passengers of an almost exact amount of delay if there is one.

REFERENCES
• https://www.gequest.com/c/flight/
• http://www.kaggle.com
• http://hive.apache.org/
• https://cwiki.apache.org/hive/gettingstarted.html
• http://www.mathworks.com/help/stats/princomp.html
• A tutorial on Principal Components Analysis
Lindsay I Smith, February 26, 2002.
• http://www.mathworks.com/help/stats/svmclassify.html

ACKNOWLEDGEMENT
• Daisy Zhe Wang. Assistant Professor, University of Florida.
• Christian Grant. Database Research Center, University of Florida.