# Project Proposal - Airline Data Analysis Using Spark Technologies

Date: April 9, 2020

To: Professor Joseph Rosen, Big Data Technology

Team Members: Bhoj Soopal (Divya), Anna Kong, Nihar Darji, Sahana Shreedhar Kulkarni

Subject: Proposal for the analysis of flight data using PySpark Tools

## Purpose Statement

Our project aims at analyzing the US domestic flights dataset using pyspark dataframes. Using Pyspark, a big data technology tool, we can therefore work with RDDs/Dataframes/Datasets in Python. Our analysis will help us find out what are the top airlines with the maximum delays/cancellations in the US. Also, our project will investigate whether there are any factors that cause flight delays/cancellations.

## Data Source

The data set is taken from the kaggle website. This consists of multiple year data from 2009 to 2018.[3]

## Proposed Tasks

1. **Data Wrangling**

Our data wrangling will involve the EDA process (Exploratory Data Analysis) as listed below:

* Clean dataset (deal with any missing values)
* Prepare dataset for analysis and modeling
* Work on visualizations

1. **Data Profiling**

Our data profiling will entail the ETL process: (Extract, Transform, Load) as listed below:

* Load dataset from AWS S3 bucket (cloud storage service)/ HDFS system
* Proceed with creating the schema on the loaded data using StructField method
* Create a structure in which the data can be read
* Produce some queries

1. **Data Modelling**

Our data modeling will include supervised learning models as listed below:

* Test these models [4], [5] to classify and predict the flight cancellations
* Perform accuracy metrics across different models
* Use PySpark and MLlib as our foundation for big data frameworks [6], [7]

1. **Documentation**

Our final deliverable will be a documented report. It will include all our data analysis, test results, any related visualizations pertaining to our research, and also our forecast modelling.

Schedule

Here is a schedule of the tasks we would complete for this project.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Tasks** | **Date of Tasks (by Weeks)** | | | | | **Owner** |
| **Data Wrangling** | x | x |  |  |  | Sahana, Nihar, Divya |
| **Data Profiling** |  | x | x |  |  | Sahana, Nihar, Divya |
| **Decide on Model, Fit, Forecast** |  |  | x | x |  | Divya, Anna |
| **Documentation** |  |  |  |  | x | Everyone |
| **Weeks** | **12** | **13** | **14** | **15** | **16** |  |
|  | **APRIL/MAY** | | | | |  |

References

**[1]** Bureau of Transportation Statistics. (2016). Airline On-Time Performance and Causes of Flight Delays. Retrieved from <https://catalog.data.gov/dataset/airline-on-time-performance-and-causes-of-flight-delays-on-time-data>

**[2]** Deshpande, V., & Arikan, M. (2011). The Impact of Airline Flight Schedules on Flight Delays. *Manufacturing & Service Operations Management, 14*, 423-440. Retrieved from <https://pubsonline.informs.org/doi/10.1287/msom.1120.0379>

**[3]** Mu, Y. (2019, August). Airline Delay and Cancellation Data, 2009 - 2018. Retrieved April 2020 from <https://www.kaggle.com/yuanyuwendymu/airline-delay-and-cancellation-data-2009-2018/data>

**[4]** Chakrabarty, Navoneel, et al. ”Flight Arrival Delay Prediction Using Gradient Boosting Classifier.” Emerging Technologies in Data Mining and Information Security. Springer, Singapore, 2019. 651-659. Retrieved from <https://www.researchgate.net/publication/327389509_Flight_Arrival_Delay_Prediction_Using_Gradient_Boosting_Classifier>

**[5]** Yi Ding ”Predicting flight delay based on multiple linear regression”, IOP Conference Series: Earth and Environmental Science. Retrieved from <https://iopscience.iop.org/article/10.1088/1755-1315/81/1/012198>

**[6]** Belcastro, L. & Marozzo, Fabrizio & Talia, Domenico & Trunfio, Paolo. (2016). Using Scalable Data Mining for Predicting Flight Delays. ACM Transactions on Intelligent Systems and Technology. 8. 10.1145/2888402. Retrieved from <https://dl.acm.org/doi/10.1145/2888402>

**[7]** Li, S. “Machine Learning with PySpark and MLlib — Solving a Binary Classification Problem”. Towards Data Science. Retrieved from

<https://towardsdatascience.com/machine-learning-with-pyspark-and-mllib-solving-a-binary-classification-problem-96396065d2aa>