W205 PROJECT: Predicting Flight Delays

Final Presentation – December 14

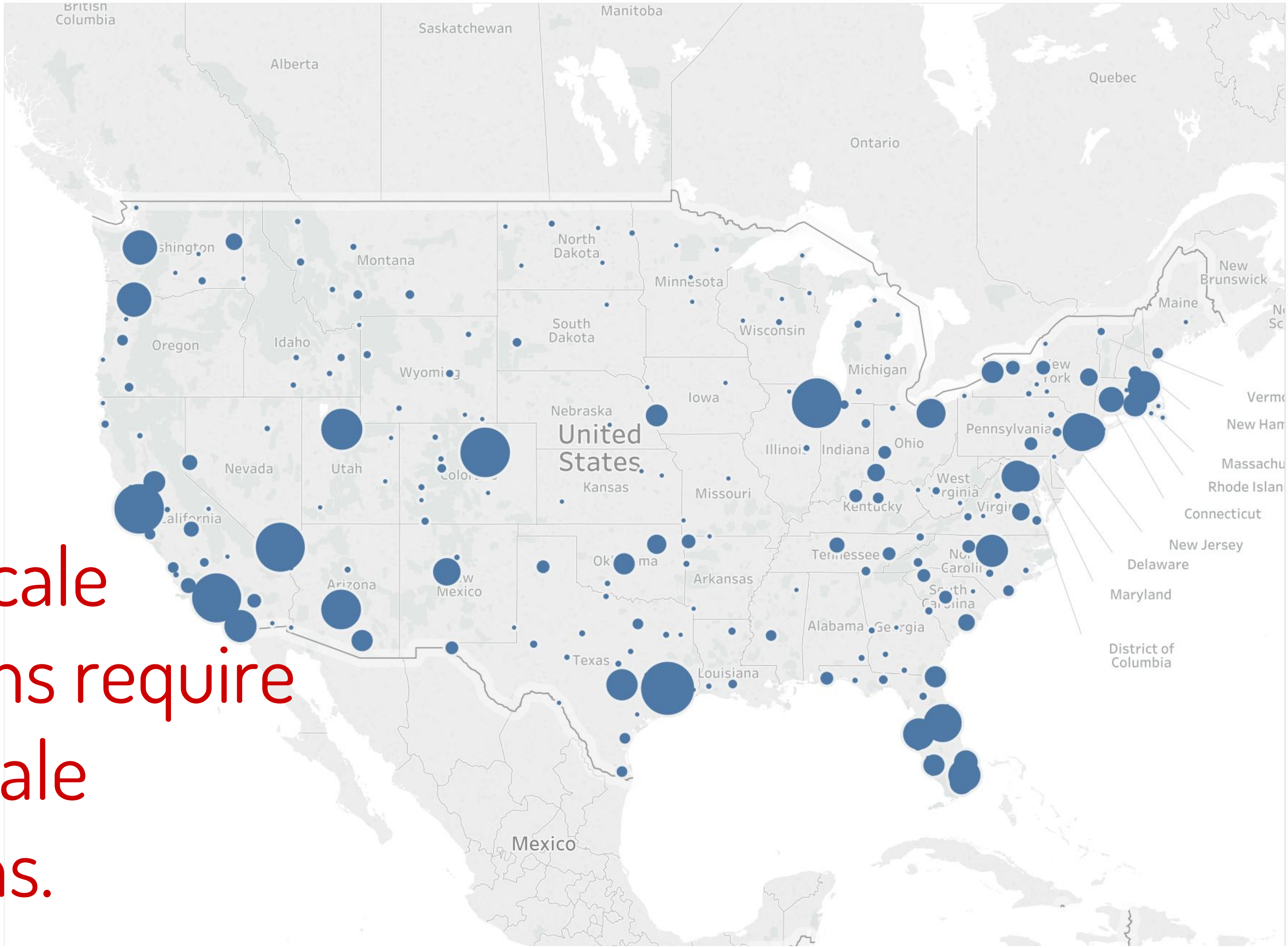
Cyprian Gascoigne, Vicki Foss, Adam Letcher

The Problem

Flight delays are costly in time to consumers and money to airlines.

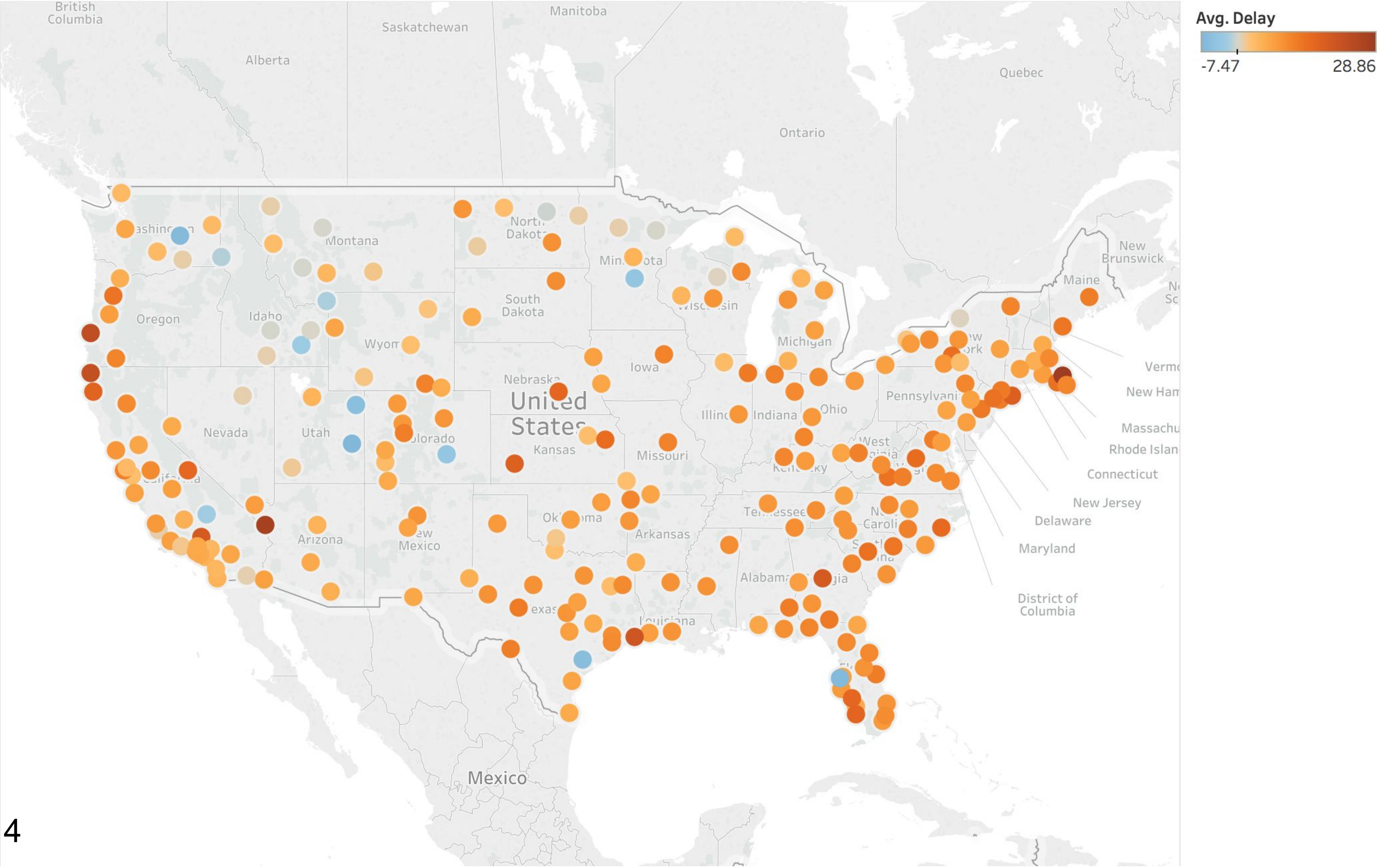
We would like to predict in real-time the probability that a given flight will be delayed.

Flight Volumes in Geographic Hubs

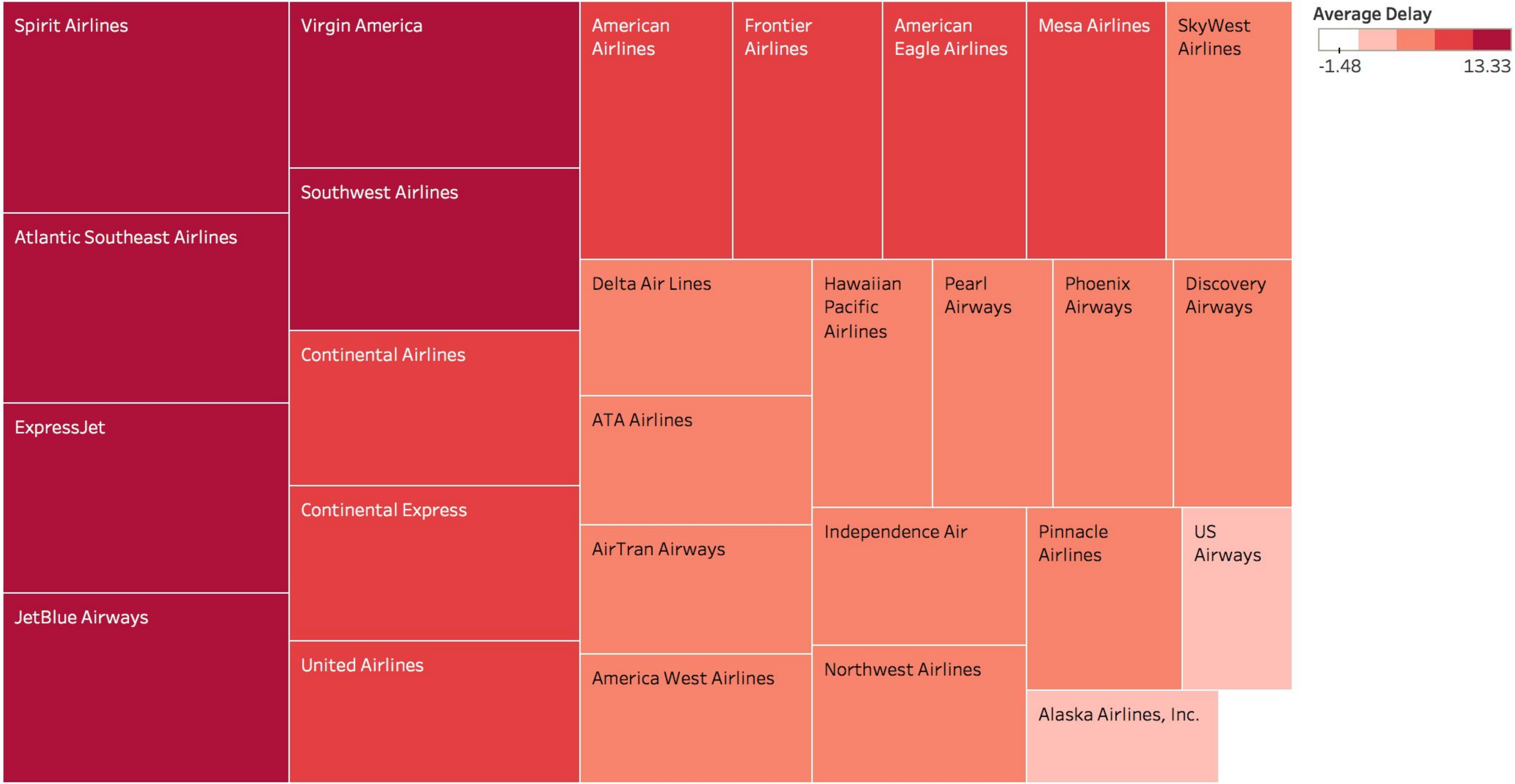


Large scale problems require large scale solutions.

Average Delay by Departure Airport



Airlines with Worst Delays



Solution Architecture

Data Sources

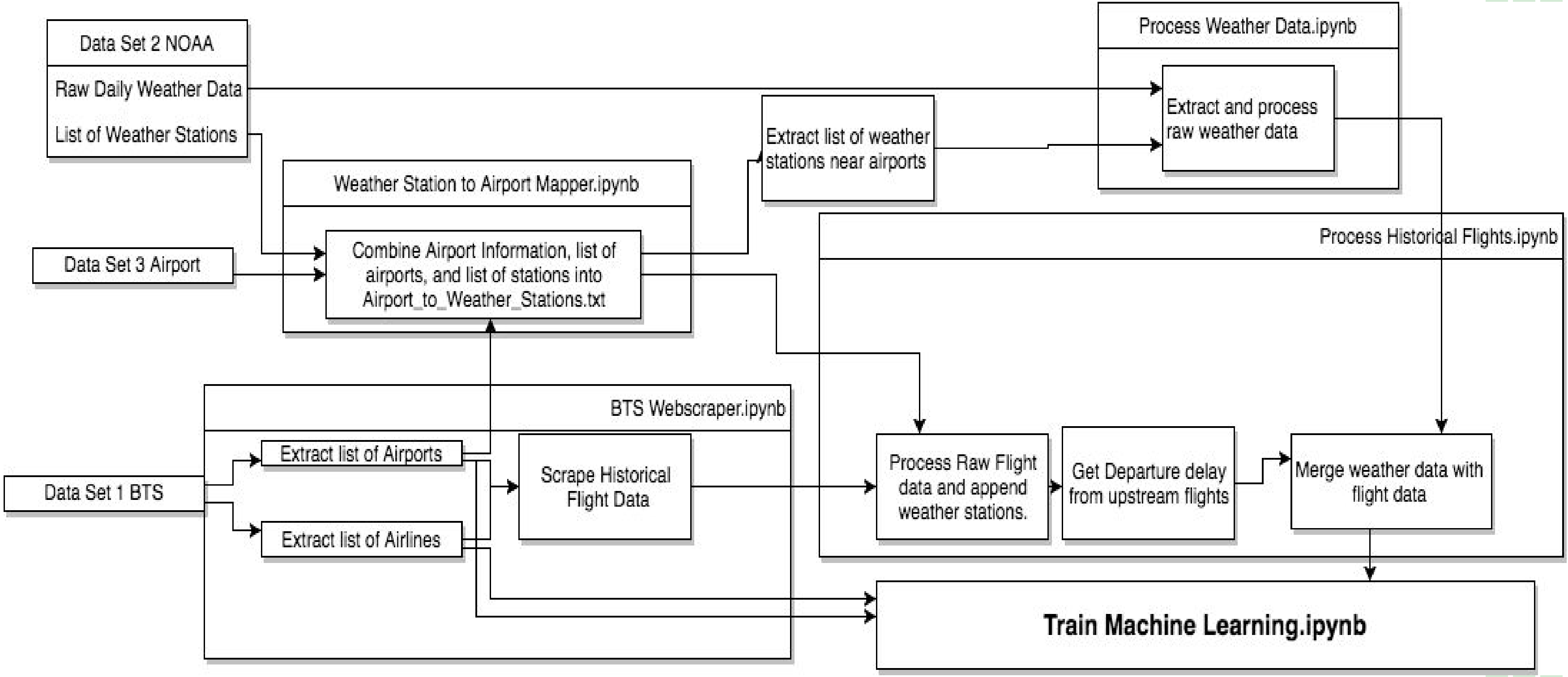
Live API Data:

- FlightStats [developer.flightstats.com]
- Wunderground [wunderground.com]

Historical Training Data:

- Flight Data: Bureau of Transportation Statistics
- Weather Data: NOAA GHCN Daily Data Set
- Airport Data: FAA Airport Data & Contact Info

Historical Data Processing


Data Set 2 NOAA
Raw Daily Weather Data
List of Weather Stations
Data Set 3 Airport
Data Set 1 BTS
Weather Station to Airport Mapper.ipynb
Combine Airport Information, list of airports, and list of stations into Airport_to_Weather_Stations.txt
Extract list of weather stations near airports
Process Weather Data.ipynb
Extract and process raw weather data
Process Historical Flights.ipynb
BTS Webscraper.ipynb
Extract list of Airports
Extract list of Airlines
Scrape Historical Flight Data
Process Raw Flight data and append weather stations.
Get Departure delay from upstream flights
Merge weather data with flight data
Train Machine Learning.ipynb

Machine Learning Prototype

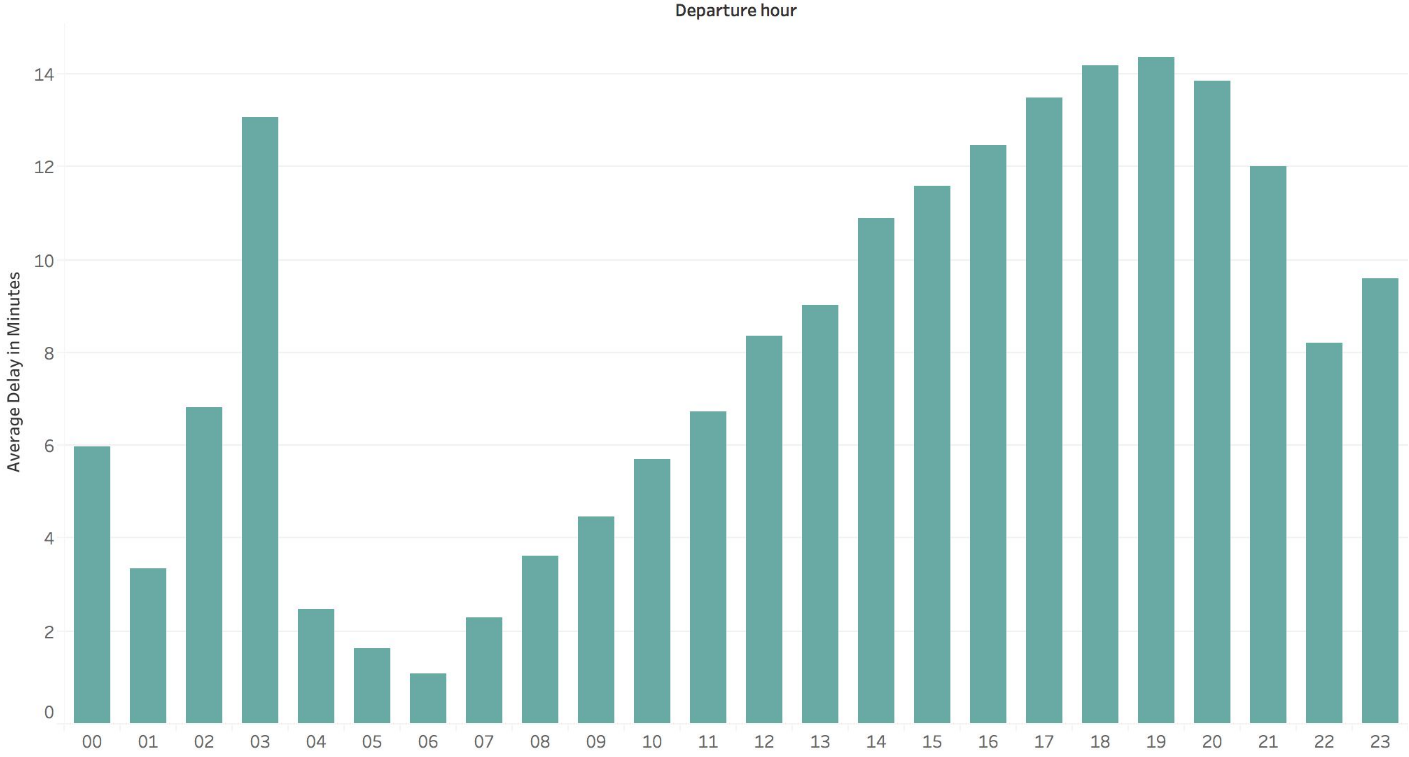
Engineering Challenges:

- ~35,000,000 rows of data
- 10-14 hour training time

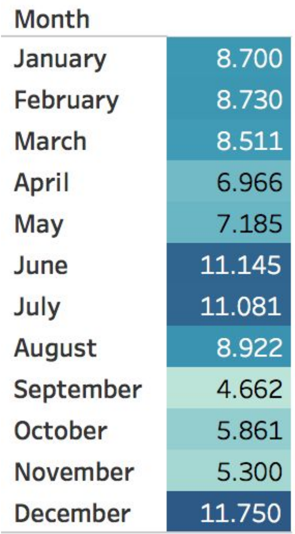
The Model:

- Stochastic Gradient Descent (70 batches)
- Label Encoding Airport, Airline
- Normalization by Range (on a sample of 1,000,000)
- Lots of room for improvement

Average delay in minutes according to scheduled departure time

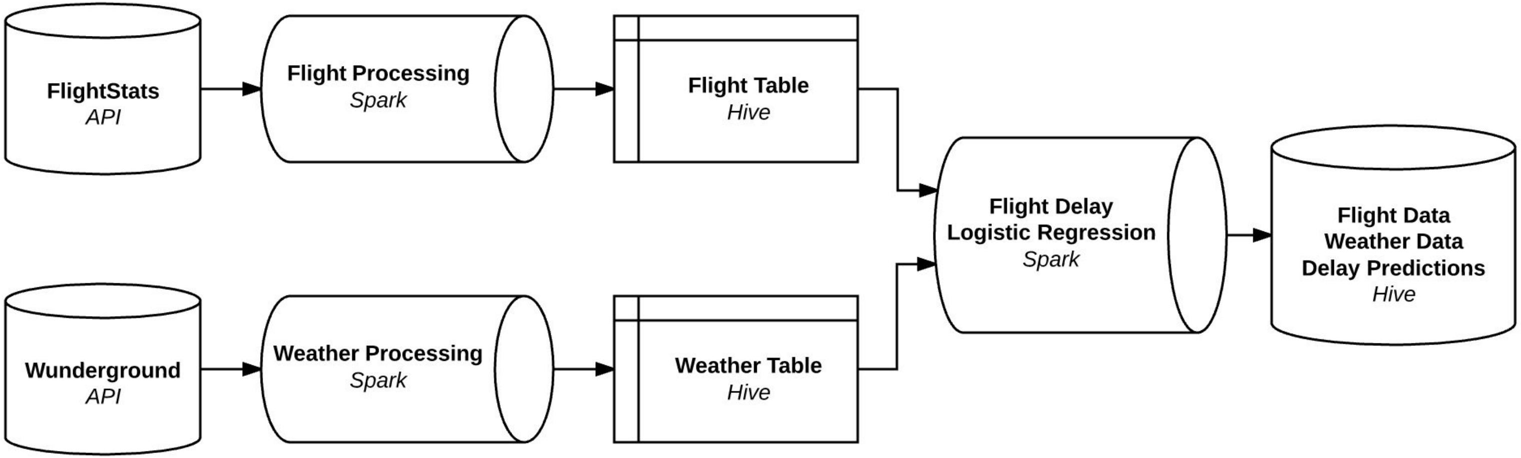


Average Delay in Minutes by Month



Month	
January	8.700
February	8.730
March	8.511
April	6.966
May	7.185
June	11.145
July	11.081
August	8.922
September	4.662
October	5.861
November	5.300
December	11.750

Current Batch Application



FlightStats
API

Flight Processing
Spark

Flight Table
Hive

Wunderground
API

Weather Processing
Spark

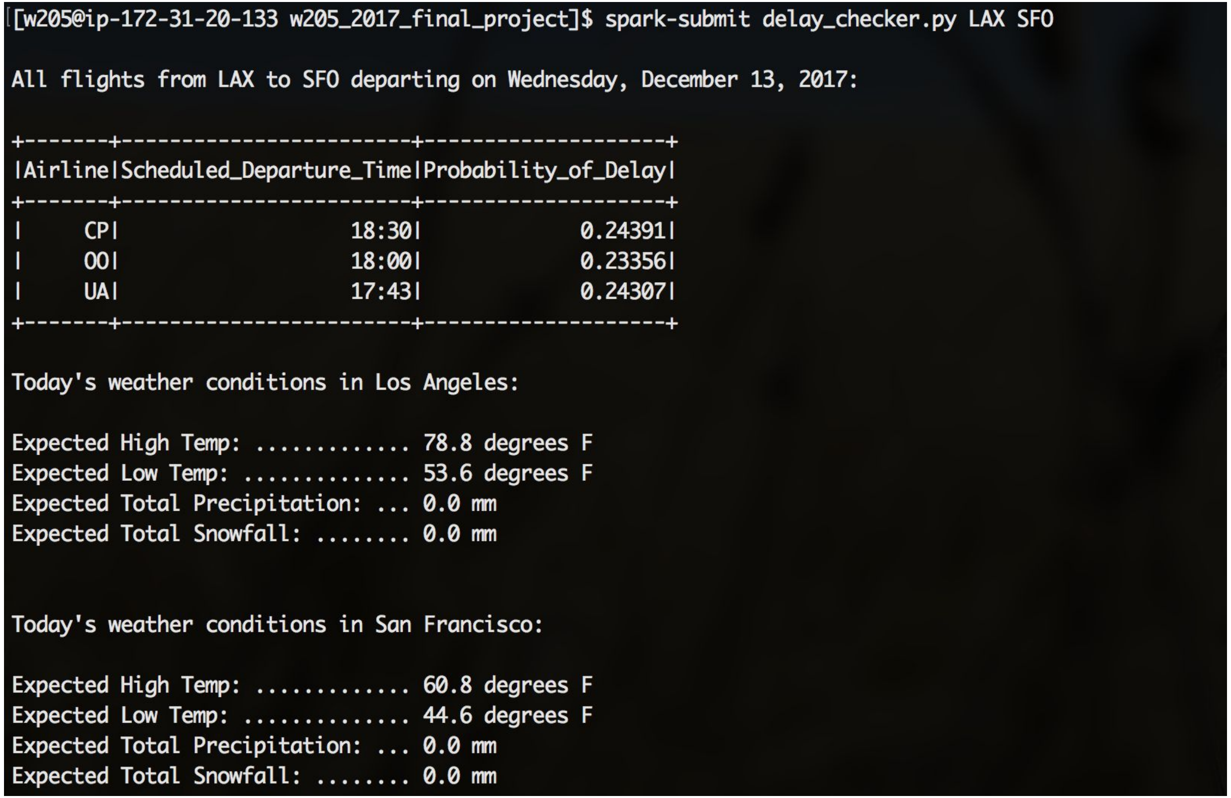
Weather Table
Hive

Flight Delay
Logistic Regression
Spark

Flight Data
Weather Data
Delay Predictions
Hive

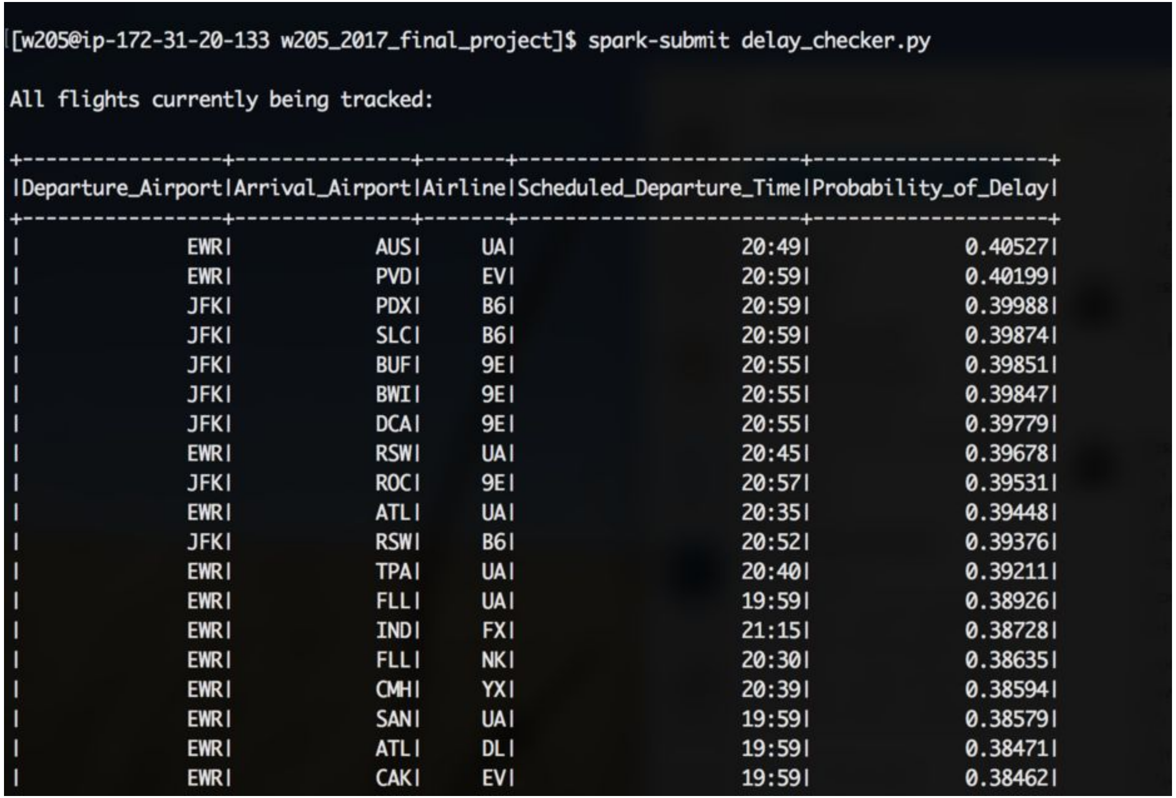
Demo / Results

Serving Script – Query



```
[w205@ip-172-31-20-133 w205_2017_final_project]$ spark-submit delay_checker.py LAX SFO

All flights from LAX to SFO departing on Wednesday, December 13, 2017:

+-------+------------------------+--------------------+
|Airline|Scheduled_Departure_Time|Probability_of_Delay|
+-------+------------------------+--------------------+
|     CP|                   18:30|             0.24391|
|     OO|                   18:00|             0.23356|
|     UA|                   17:43|             0.24307|
+-------+------------------------+--------------------+

Today's weather conditions in Los Angeles:

Expected High Temp: ............. 78.8 degrees F
Expected Low Temp: .............. 53.6 degrees F
Expected Total Precipitation: ... 0.0 mm
Expected Total Snowfall: ........ 0.0 mm


Today's weather conditions in San Francisco:

Expected High Temp: ............. 60.8 degrees F
Expected Low Temp: .............. 44.6 degrees F
Expected Total Precipitation: ... 0.0 mm
Expected Total Snowfall: ........ 0.0 mm
```

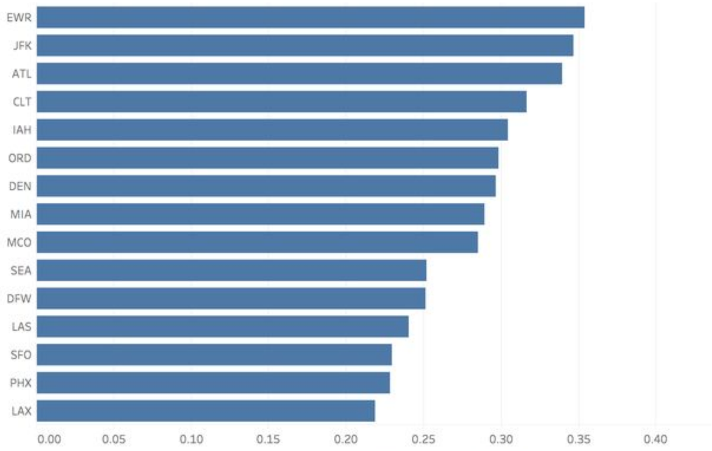
Serving Script – All



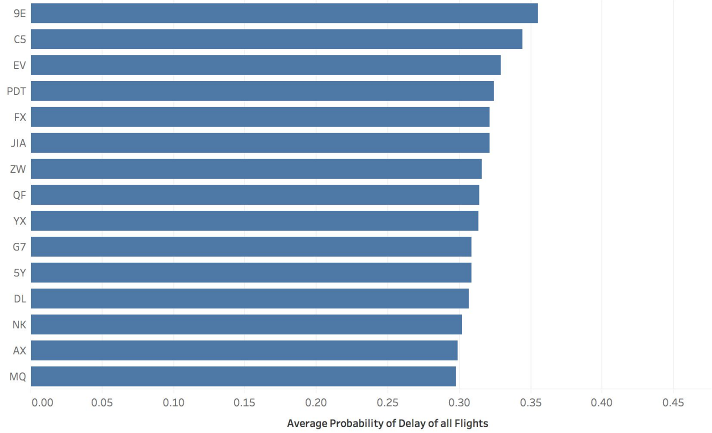
```
[[w205@ip-172-31-20-133 w205_2017_final_project]$ spark-submit delay_checker.py

All flights currently being tracked:

+-----------------+---------------+-------+------------------------+--------------------+
|Departure_Airport|Arrival_Airport|Airline|Scheduled_Departure_Time|Probability_of_Delay|
+-----------------+---------------+-------+------------------------+--------------------+
|              EWR|            AUS|     UA|                   20:49|             0.40527|
|              EWR|            PVD|     EV|                   20:59|             0.40199|
|              JFK|            PDX|     B6|                   20:59|             0.39988|
|              JFK|            SLC|     B6|                   20:59|             0.39874|
|              JFK|            BUF|     9E|                   20:55|             0.39851|
|              JFK|            BWI|     9E|                   20:55|             0.39847|
|              JFK|            DCA|     9E|                   20:55|             0.39779|
|              EWR|            RSW|     UA|                   20:45|             0.39678|
|              JFK|            ROC|     9E|                   20:57|             0.39531|
|              EWR|            ATL|     UA|                   20:35|             0.39448|
|              JFK|            RSW|     B6|                   20:52|             0.39376|
|              EWR|            TPA|     UA|                   20:40|             0.39211|
|              EWR|            FLL|     UA|                   19:59|             0.38926|
|              EWR|            IND|     FX|                   21:15|             0.38728|
|              EWR|            FLL|     NK|                   20:30|             0.38635|
|              EWR|            CMH|     YX|                   20:39|             0.38594|
|              EWR|            SAN|     UA|                   19:59|             0.38579|
|              EWR|            ATL|     DL|                   19:59|             0.38471|
|              EWR|            CAK|     EV|                   19:59|             0.38462|
```

Tableau Interface with Hive Results

Probability of Delay for Flights according to Departure Airport (December 13)



Probability of Delay for Current Flights according to Airline (Top 15)



Streaming Demonstration Video

Potential Enhancements and Scale-out

Enhancing and Scaling the Application

Stream Processing

- API continuous query to provide constant update
- Spark streaming to process data flows and update tables automatically

Machine Learning

- Fuzzy data methods
- Implement upstream flights
- Hyperparameter tuning
- Testing other models (RF)
- Binarize Inputs

Scale-out

- Spark streaming
- Restructure scripts
- Hive
- Scalable Visualizations
- Distributed Machine Learning