Team: Munieshwar (Kevin) Ramdass

Om Narayan

Professor Juan Rodriguez

CS-GY 9223 – Programming in Big Data

10th May 2017

Aviation Analysis

For Airports, For People

1 Summary

We analyzed air traffic data from **Air Travel Consumer Reports**. We hope to provide a means to which airport traffic can be made scalable and manageable. To do so we analyzed plane departure and arrival on an hourly, daily, weekly, monthly, and annual basis to give a high level view of which seasons/periods are more busy or less busy. The analysis can be used to provide a smooth scalability for air traffic in the future by accommodating more flight intake and outgoing traffic with a higher number of destinations. The hope is that travelers can also view this data to plan their next flight or enjoy some fun facts.

2 Technology Used

Spark 2
Python 2.7 (standalone and pySpark)
Spark SQL
Elasticsearch
Hadooop (for testing purposes - not for submission)
Pig Latin (for testing purposes - not for submission)

3 Using this Data

Our hope is that people and airports can use this data to better plan their day. Taking into considerations which days are the busiest at a given airport (in terms of how many flights are expected to land verses depart) is good to know for airports to raise the employee count for that day. This can streamline airport processes. People might be interested to learn what flights to avoid by looking at the delay times for certain carriers. Taxi times to arrive and depart is also a factor in determining how cluttered an airport can be. Taking into consideration which days and month have the longest taxi delays can help airports decide how to orient traffic to service arriving passengers quickly.

4 Architecture

We used NYU HPC to run our analysis. NYU DUMBO has Spark. NYU Babar has HDFS. We SSH into DUMBO and ran queries via Spark. Results saved on HDFS were then ported to an Elasticsearch instance. From here we were able to generate visualizations.

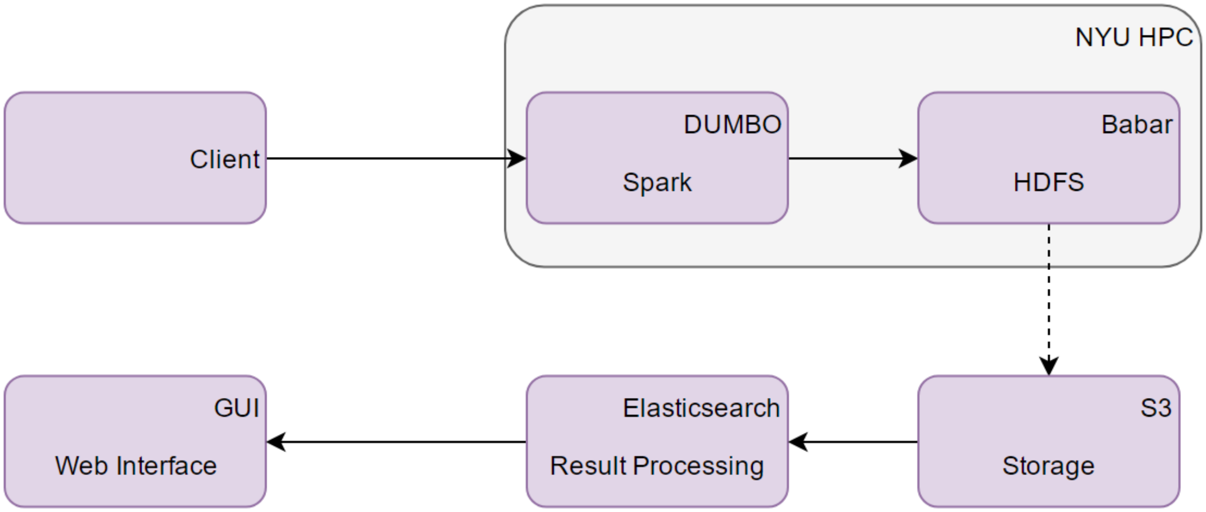


Figure 3.1: Architecture

5 Queries Done

The following are queries in English that we ran on Spark. The actual queries were written using pySpark and took advantage of SQL statements (Spark SQL). This can be found in **driver.py** and **more_queries.py**.

For queries 1 to 3, we wrote statements that found the count of ConsumerReport between source and destination for each Year, Month, DayofMonth, DayOfWeek.

Query 4 finds the total number of ConsumerReport delayed or diverted during the year.

Query 5 finds total number of ConsumerReport delayed or diverted during the year on monthly basis.

Query 6 finds total number of ConsumerReport delayed or diverted during the year for each Airlines.

Query 7 finds total number of ConsumerReport diverted during the year for source and destination pair.

Query 8 finds the total number of ConsumerReport cancelled during the year for source and destination pair.

Queries 9 to 12 shows seasonal trends or counts of flights. The idea behind these queries were to find any major differences within destinations that flights go to which is found using the next set of queries.

Queries 13 to 16 describes the destinations of flights per season in hopes of explaining the counts found in the last set of queries. We wanted to know if the rank of count between airports changed drastically in any way.

Queries 17 and 18 details the average arrival and departure times respectively for each airlines.

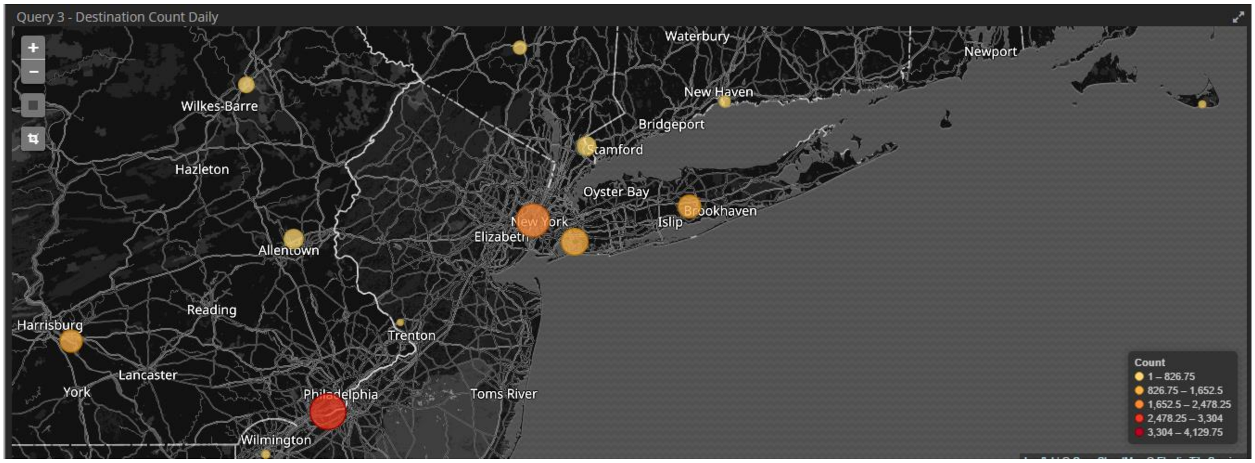
Queries 19 to 22 were redundancies of queries 13 to 16 with the exception that column names were added. The was merely done for experimentation.

Queries 23 to 26 are similar to queries 13 to 16. Instead, they show the sources of flights per season.

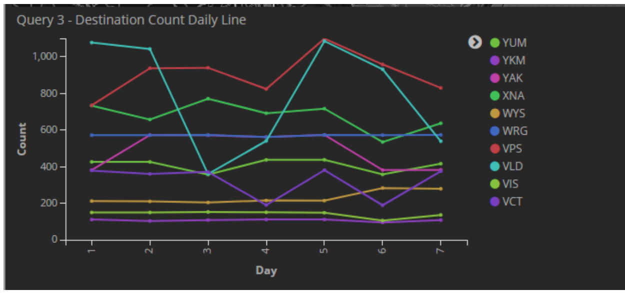
Queries 27 to 30 were redundancies of queries 23 to 26 with the exception that column names were added. The was merely done for experimentation.

6 Results and Discoveries

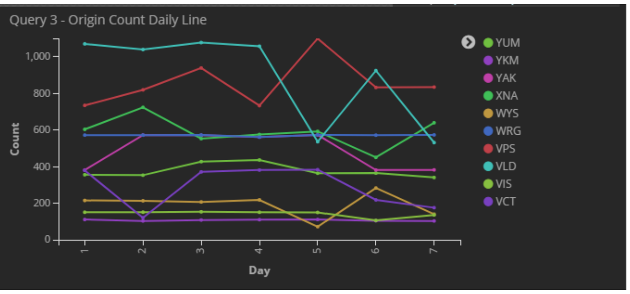
We thought that winter would have the most flights. It turns out to be the opposite. Summer has the most flights while winter has the least. We realized that it is best to avoid flight number 9024 as it has an average delay time of 1392 minutes. This could mean that it has been diverted a lot and the arrival time added up. Other critical data shows that some airports have more outgoing flights than incoming flights per day and vice versa. This is interesting in that it helps assess how busy airports are. More incoming flights than outgoing flights per day means that airports will be busier.



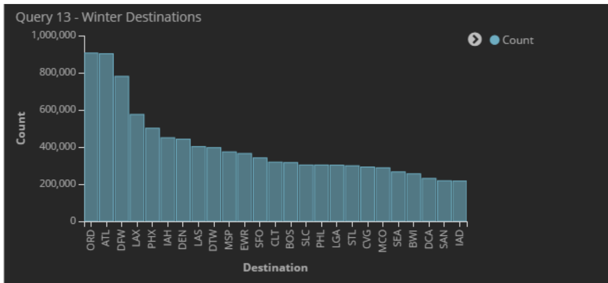
Query 3 - Destination Count Daily Line



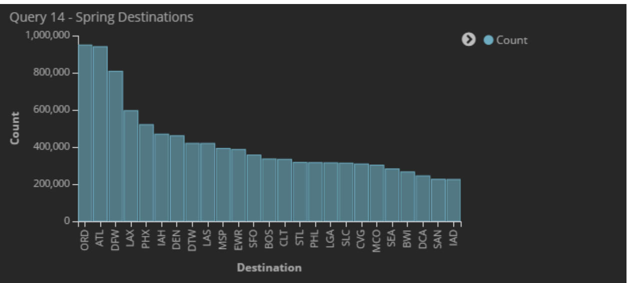
Query 3 - Origin Count Daily Line



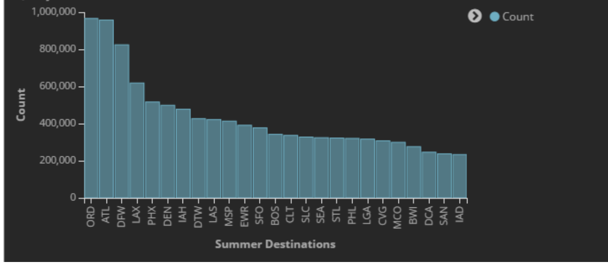
Query 13 - Winter Destinations



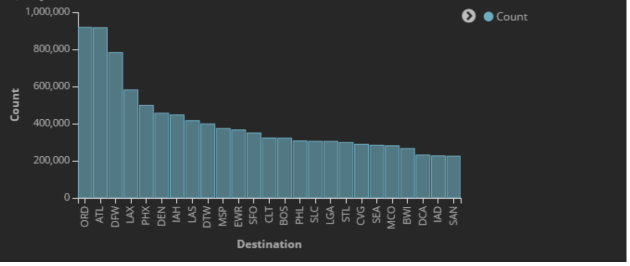
Query 14 - Spring Destinations



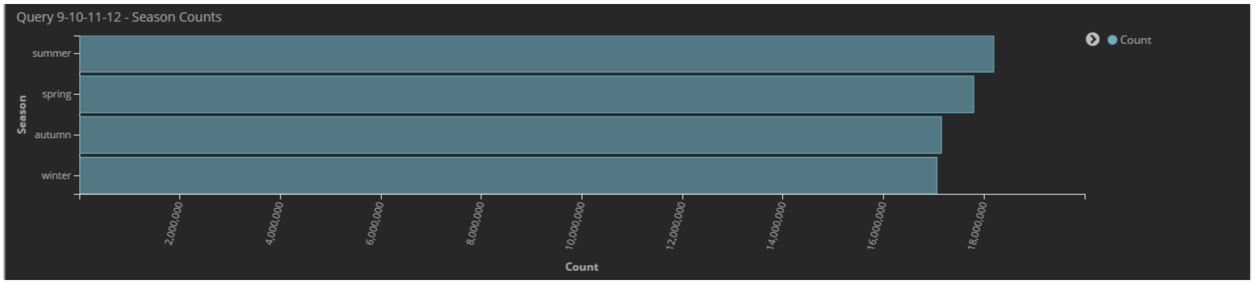
Query 15 - Summer Destinations



Query 16 - Autumn Destinations



Query 9-10-11-12 - Season Counts

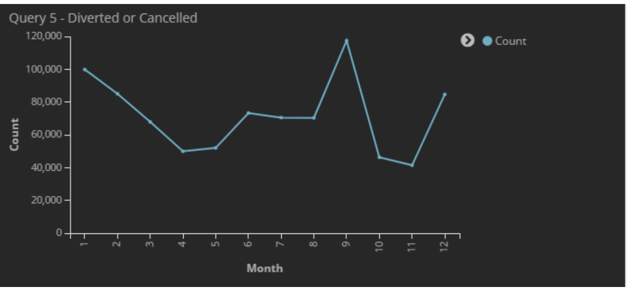


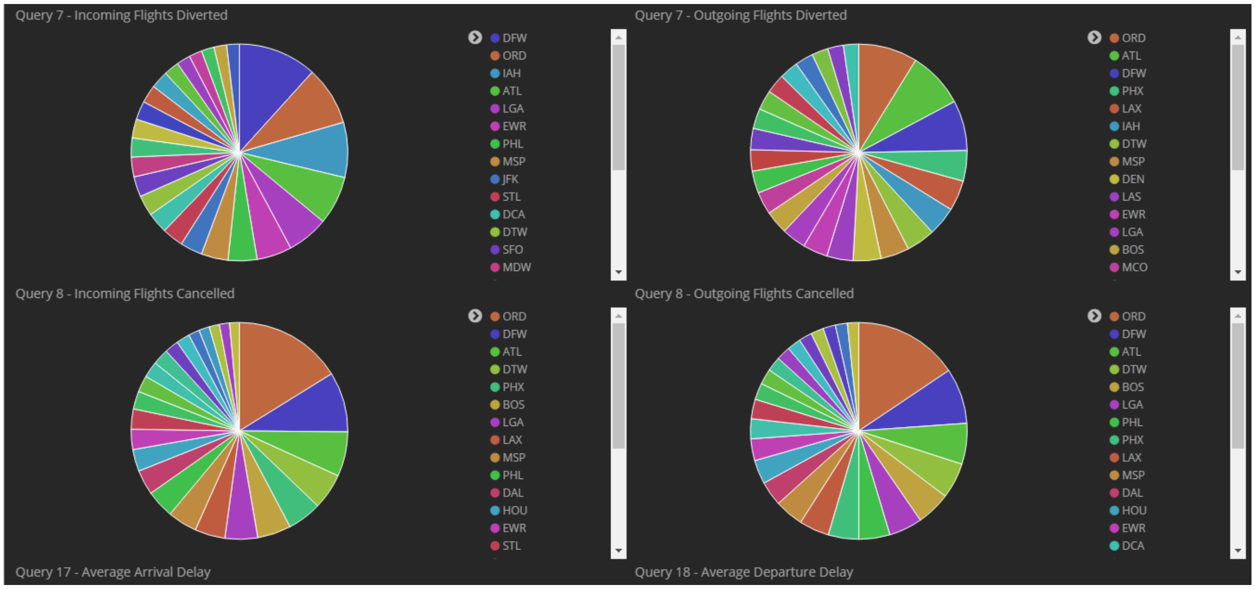
Query 4 - Cancelled or Diverted

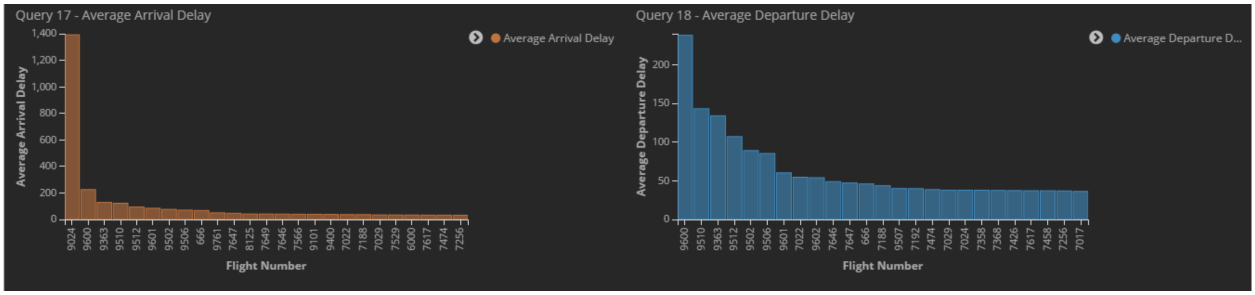
1,717,781

Count

Query 5 - Diverted or Cancelled




Query 7 - Incoming Flights Diverted
DFW
ORD
IAH
ATL
LGA
EWR
PHL
MSP
JFK
STL
DCA
DTW
SFO
MDW
Query 7 - Outgoing Flights Diverted
ORD
ATL
DFW
PHX
LAX
IAH
DTW
MSP
DEN
LAS
EWR
LGA
BOS
MCO
Query 8 - Incoming Flights Cancelled
ORD
DFW
ATL
DTW
PHX
BOS
LGA
LAX
MSP
PHL
DAL
HOU
EWR
STL
Query 8 - Outgoing Flights Cancelled
ORD
DFW
ATL
DTW
BOS
LGA
PHL
PHX
LAX
MSP
DAL
HOU
EWR
DCA
Query 17 - Average Arrival Delay
Query 18 - Average Departure Delay



Link to our website: http://aviation-traffic-analysis.ehmknmds2s.us-east-1.elasticbeanstalk.com/

Link to our GitHub repository: https://github.com/mramdass/Aviation_Traffic_Analysis

Link to the Data: http://stat-computing.org/dataexpo/2009/the-data.html