**INFO 7250 – Engineering of Big Data**

**PROJECT PROPOSAL**

I am using the data on **Airline On-Time Statistics and Delay Causes from**

<http://stat-computing.org/dataexpo/2009/the-data.html>

This is dataset containing information about airline schedule with following columns:

**Variable descriptions**

|  |  |  |
| --- | --- | --- |
|  | **Name** | **Description** |
| 1 | Year | 1987-2008 |
| 2 | Month | 1-12 |
| 3 | DayofMonth | 1-31 |
| 4 | DayOfWeek | 1 (Monday) - 7 (Sunday) |
| 5 | DepTime | actual departure time (local, hhmm) |
| 6 | CRSDepTime | scheduled departure time (local, hhmm) |
| 7 | ArrTime | actual arrival time (local, hhmm) |
| 8 | CRSArrTime | scheduled arrival time (local, hhmm) |
| 9 | UniqueCarrier | [unique carrier code](http://stat-computing.org/dataexpo/2009/supplemental-data.html) |
| 10 | FlightNum | flight number |
| 11 | TailNum | plane tail number |
| 12 | ActualElapsedTime | in minutes |
| 13 | CRSElapsedTime | in minutes |
| 14 | AirTime | in minutes |
| 15 | ArrDelay | arrival delay, in minutes |
| 16 | DepDelay | departure delay, in minutes |
| 17 | Origin | origin [IATA airport code](http://stat-computing.org/dataexpo/2009/supplemental-data.html) |
| 18 | Dest | destination [IATA airport code](http://stat-computing.org/dataexpo/2009/supplemental-data.html) |
| 19 | Distance | in miles |
| 20 | TaxiIn | taxi in time, in minutes |
| 21 | TaxiOut | taxi out time in minutes |
| 22 | Cancelled | was the flight cancelled? |
| 23 | CancellationCode | reason for cancellation (A = carrier, B = weather, C = NAS, D = security) |
| 24 | Diverted | 1 = yes, 0 = no |
| 25 | CarrierDelay | in minutes |
| 26 | WeatherDelay | in minutes |
| 27 | NASDelay | in minutes |
| 28 | SecurityDelay | in minutes |
| 29 | LateAircraftDelay | in minutes |

The reason of selections this data set is that it has many numbers of columns which will enable me to use various MapReduce algorithms studies in the course for different types of analysis.

Also, the data is evenly segregated in yearly basis. So, in case If I can am unable to load complete data in my computer then too I can do the same analysis on small portion of same data more easily.

**In this project I will try to answer following questions :-**

1. Which month, time or day of week contributed in maximum delay in airline departure and/or arrivals ?
2. At what time during the day the airlines are most busy?
3. What were various causes of delay ?
4. Depending on departure and arrival time, which destination is best efficient from which starting city?

Apart from above analysis, I will try to do more analysis using Apache Hive and Apache Pig.