**INFO 7250 – Engineering of Big Data Systems**
**Project Proposal – Initial Plan**

I have selected the flight arrival data from the list of datasets professor provided.
The data can be found from the below link:
http://stat-computing.org/dataexpo/2009/the-data.html

The main reason for selecting the data is that it is vast and spread over for several years. So, a wide range of analysis can be done considering various parameters.
The data has following attributes/columns:

| | 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 |
| 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 |
| 18 | Dest | destination IATA airport code |
| 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 |

As of now, I am planning to do the following analysis on data which can be modified later to add more type of analyses:

1. Which year has maximum delay for Arrival or Departure? (The same analysis can be done to find for which month or day)
2. Which flight was cancelled for most times and what are the main reasons for it?
3. Which is the peak day in a month in which the airlines are busy? (The same analysis can be done to find average peak time in a day)
4. Which origin airport has least departure delay time and which destination airport has least arrival delay time? (also vice versa can be done for highest time or we can also find top 5 airports instead of the most in any case)
5. What is the average number of flights diverted per year?
6. How many flights arrived on time in a year/month/day?

I will try to solve these problems using MapReduce at present. Later, I will try to do more analysis using Pig, Hive, Mahout etc