## Working with an Airline dataset

* This exercise is inspired by <http://www.datadr.org/doc/airline.html>
* Full information on datasets (optional datasets), and general documentation available here: <http://stat-computing.org/dataexpo/2009/>

Before we start, here's a description of the dataset "schema". We will work on data that can be downloaded from here: <http://stat-computing.org/dataexpo/2009/the-data.html>

Note that there is a single CSV file per year, hence the first field below is somehow redundant, although you can imagine to concatenate all files and work on them as a whole (which by the way would make sense when using Hadoop MapReduce / Pig). In summary, there are 29 fields which provide enough information to build Pig scripts that cover Queries 1-5. For the advanced analysis subsection, you need other data, which can be downloaded from the links below.

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

Other sources of data come from here: <http://stat-computing.org/dataexpo/2009/supplemental-data.html>. Precisely, we are interested in:

* Airport IATA Codes to City names and Coordinates mapping: <http://stat-computing.org/dataexpo/2009/airports.csv>
* Carrier codes to Full name mapping: <http://stat-computing.org/dataexpo/2009/carriers.csv>
* Information about individual planes: <http://stat-computing.org/dataexpo/2009/plane-data.csv>
* Weather information: <http://www.wunderground.com/weather/api/>. You can subscribe for free to the developers API and obtain (at a limited rate) hystorical weather information in many different formats. Also, to get an idea of the kind of information is available, you can use this link: <http://www.wunderground.com/history/>

### Query 1: Top 20 cities by total volume of flights

What are the busiest cities by total flight traffic. JFK will feature, but what are the others? For each airport code compute the number of inbound, outbound and all flights. Variation on the theme: compute the above by day, week, month, and over the years.

### Query 2: Carrier Popularity

Some carriers come and go, others demonstrate regular growth. Compute the (log base 10) volume -- total flights -- over each year, by carrier. The carriers are ranked by their median volume (over the 10 year span).

### Query 3: Proportion of Flights Delayed

A flight is delayed if the delay is greater than 15 minutes. Compute the fraction of delayed flights per different time granularities (hour, day, week, month, year).

### Query 4: Carrier Delays

Is there a difference in carrier delays? Compute the proportion of delayed flights by carrier, ranked by carrier, at different time granularities (hour, day, week, month year). Again, a flight is delayed if the delay is greater than 15 minutes.

### Query 5: Busy Routes

Which are busy the routes? A simple first approach is to create a frequency table for the unordered pair (i,j) where i and j are distinct airport codes.

### Advanced analyses

* When is the best time of day/day of week/time of year to fly to minimise delays?
* Do older planes suffer more delays?
* How does the number of people flying between different locations change over time?
* How well does weather predict plane delays?
* Can you detect cascading failures as delays in one airport create delays in others? Are there critical links in the system?