Case Study on Flight Data

Some insights on the U.S. Airline data using Apache Pig

Understanding Data

The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics (BTS) tracks the on-time performance of domestic flights operated by large air carriers. Summary information on the number of on-time, delayed, cancelled, and diverted flights appears in DOT's monthly Air Travel Consumer Report, published about 30 days after the month's end.

The data format is comma separated values. These are 2 different datasets, i.e., delayedflights.csv and airports.csv. Let us understand one at a time.

1. delayedflights.csv
2. airports.csv

There are 29 columns in delayedflights.csv dataset. Please check flight_description file for details about schema/fields.

For airports.csv

- iata: the international airport abbreviation code
- name of the airport
- city and country in which airport is located.
- lat and long: the latitude and longitude of the airport

Some exploration ideas with Pig:

1. Find out the top 5 most visited destinations.
2. Which month has seen the most number of cancellations due to bad weather?

3. Top ten origins with the highest AVG departure delay

4. Which route (origin & destination) has seen the maximum diversion?