SparkFlight4

November 22, 2015

In [2]:
```
from operator import add
data = sc.textFile("../big_data_eszkozok/hazi/input/2008.csv")
header = data.first()
rows = data.filter(lambda line: line != header)\ #header eldobás
.map(lambda line: line.split(","))\     #mezők kinyerése
.filter(lambda line: len(line)>1)\      #üres sorok eldobása
.map(lambda line: (line[16],line[15]))\ #releváns mezők kiválasztása
.filter(lambda line: line[1] != "NA")\  #nem teljes sorok kihagyása
.map(lambda line: (line[0], int(line[1])))\ #mező int konverziója
.filter(lambda line: line[1] > 0) \     #szűrés későn indulásra
.map( lambda line : (line[0], 1)) \     #későn indulások számolása
.reduceByKey(add) \
.takeOrdered(10, key=lambda x: -x[1])   #legtöbbet késő 10 kiválasztás
#rows = [i[0] for i in rows]
rows
```

Out[2]:
```
[(u'ATL', 175017),
 (u'ORD', 159427),
 (u'DFW', 127749),
 (u'DEN', 104414),
 (u'LAX', 87258),
 (u'IAH', 87139),
 (u'PHX', 82915),
 (u'LAS', 76240),
 (u'EWR', 69612),
 (u'DTW', 59837)]
```