A5

Abhishek Ravi Chandran, Chinmayee Vaidya

March 14, 2016

Design

In this assignment we try to find the number of connecting flights for each carrier for every year.

In the mapper class we use a custom key where we write every record twice differentiating them by a flag(type). 0 for incoming flights and 1 for outgoing flights.

The logic for handling overnight flights is already present in the method where we set values to our object after parsing it.

We sort the data to make sure that all incoming flights are sent to the reducer first, followed by all the outgoing flights. This is achieved by having a custom key, custom sort comparator and a custom group comparator. By using the group comparator we ensure that each carrier and year are chosen as the natural key to send all records grouped with this key to a single reducer.

In the reducer, we put all the incoming flights into a map with the key as city(destination), so all flights incoming to city 'a' will be under the key 'a'. The records are stored in a TreeSet which will sort all the flights stored according to their scheduled arrival time.

When we start getting outgoing flights, for each record we fetch all flights incoming to the same city as the origin city of this record, by getting the records from the map. Since the values are sorted, we keep checking for connections until we reach a record where the time difference is more than 360 minutes, meaning any flight after this will not form a connection as it is outside the time window. This ensures that not all flights are not compared reducing the number of comparisons.

Conditions for a connection: A Connections is made when A.destination = B.Origin && [(B.scheduledDeparture - A.ScheduledArrival) >=30 mins && <=6hrs]

A connection is said to be bad if flight A is cancelled or(B.actualDeparture - A.actualArrival) < 30 mins

Performance: The program runs for ~22 minutes in a 5 m3.xlarge cluster configuration for 2 years data.

Output:

```
##  Carrier Year Connection Missed Connection
##       AA 2013   23346062            689448
##       AA 2014   23806522            849930
##       AA 2015    1737811             62572
##       UA 2013   13596609            486001
##       UA 2014   13192106            524387
##       UA 2015   25484847            909645
##       WN 2013   69895212           1966131
##       WN 2014   54030502           1740104
##       WN 2015   15486373            581425
##       FL 2013   29902611            920100
##       FL 2014   24697492            880688
##       DL 2013   47181862           1550486
##       DL 2014   52753867           1635183
##       DL 2015   27177304            785440
```

```
##       US 2013   42418913           1283276
##       US 2014   65101656           2010867
##       US 2015   49198284           1606878
##       B6 2013    2829709            134690
##       B6 2014    2885561            123047
##       B6 2015     224252              9998
##       F9 2013    2073521             96489
##       F9 2014    1285683             71906
##       F9 2015   47220320           1553496
##       HA 2013   48260592           1577915
##       HA 2014    2371900            101176
##       HA 2015    2164078             99552
##       OO 2013   28054587           1381766
##       OO 2014   13733918            671696
##       OO 2015    3177842            149089
##       EV 2013   19557578            975703
##       EV 2014   15403386            809565
##       EV 2015     962839             46605
##       YV 2013   29422962           1447097
##       VX 2013   19911200            990640
##       VX 2014   12181116            591394
##       VX 2015    9344752            492777
##       MQ 2013   11821823            577159
##       MQ 2014    9284937            490479
##       MQ 2015    6881709            275467
##       NK 2015    9319411            492027
##       9E 2013    6197878            237587
##       AS 2013    1664334             45499
##       AS 2014    1789073             54049
##       AS 2015     134977              4862
```

Assumptions:

We have not taken care of Year rollover as the year is part of the key. We assume that the number of flights will be minimal for the end of the year and should not change the result much.