A6 Report

Abhishek Ravi Chandran, Chinmayee Vaidya

March 7, 2016

High Level Design:

We used **Weka machine learning** library in this assignment. We have two jobs pipelined to process the data.

The first job reads the historical input files and separates them with the keys carrier and quarter.

We perform data cleanup by dropping all corrupt records, and flights which are cancelled. We chose the following attributes for generating the random forest classifier.

Attributes

- month(1-12)
- day of the week(1-7)
- scheduled arrival time(time/10 – (0-24))
- scheduled departure time(time/10 – (0-24))

We chose to have a maximum of 54 levels

- scheduled elapsed time(time/100 – (0-53))
- distance of the journey(distance/100 – (0-53))

We chose to keep only airports with highest traffic

- origin(ATL,ORD,DFW,LAX,DEN,IAH,PHX,SFO,CLT,DTW,MSP,LAS,MCO,EWR,JFK,LGA,BOS,SLC,SEA, BWI,MIA,MDW,PHL,SAN,FLL,TPA,DCA,IAD,HOU,??)
- destination(ATL,ORD,DFW,LAX,DEN,IAH,PHX,SFO,CLT,DTW,MSP,LAS,MCO,EWR,JFK,LGA,BOS,SLC,SEA, BWI,MIA,MDW,PHL,SAN,FLL,TPA,DCA,IAD,HOU,??)
- holiday(0,1)[1-journey is near a holiday(we kept the range as a 2 days)]

Only the data needed is written into a arff file with the key as the name and uploaded into s3

Job 2 in the pipeline has the test data as the input. The same keys as above are chosen and passed to the reducer. In the reducer the arff file is downloaded with the same name as the key. This is used to create the training dataset. We form a random forest with 10 trees. The training data is converted into the training dataset. each record is classified to get the prediction for each flight, which is written to the output of the reducer.

We download the output files and run a java program to compare our results with the validate file given to form the confusion matrix and calculate the accuracy as shown below:

Confusion matrix

```
##        TRUE     FALSE
## TRUE   1093088  775127
## FALSE  959460  899239
##
## accuracy:0.5345782059902643
```

Performance:

The program takes ~9 minutes in a configuration of 4 m3.xlarge machines