OTP Flight Ticket Price Analysis with Hadoop

Jun Cai

January 31, 2016

This document describes an implementation of Mapreduce instance with Hadoop on the OTP flight ticket price dataset. The purpose of the application is to calculate the average ticket price for each month for each carrier in the given data. The application can run in two modes: local pseudo-distributed mode and cluster mode using AWS EMR instance.

Implementation

Input Data Processing

The format of the input data is compressed csv files. These gzipped files are store in the 'input' folder residing in HDFS under pseudo-distributed mode and in S3 bucket under cluster mode. Since the HDFS can automatically handle the decompressing of the gzipped file, there is no need to explicitly write code for that. The unzipped csv files will be the input of the Mapreduce. They will be read as a text file. Each line is parsed as a record using a simplified csv format parser.

Mapping

The mapper will map each record in the given data files to a key value pair, where the key is the carrier's unique identifier and the value is a string contains the date of the flight and the average ticket price of that flight. In this way we can calculate average ticket price for each month at the later reduce step. Records with invalid format will be mapped to a special key value pair to keep tracking the number of bad records. The key and value will all be string "INVALID" in this case.

Reducing

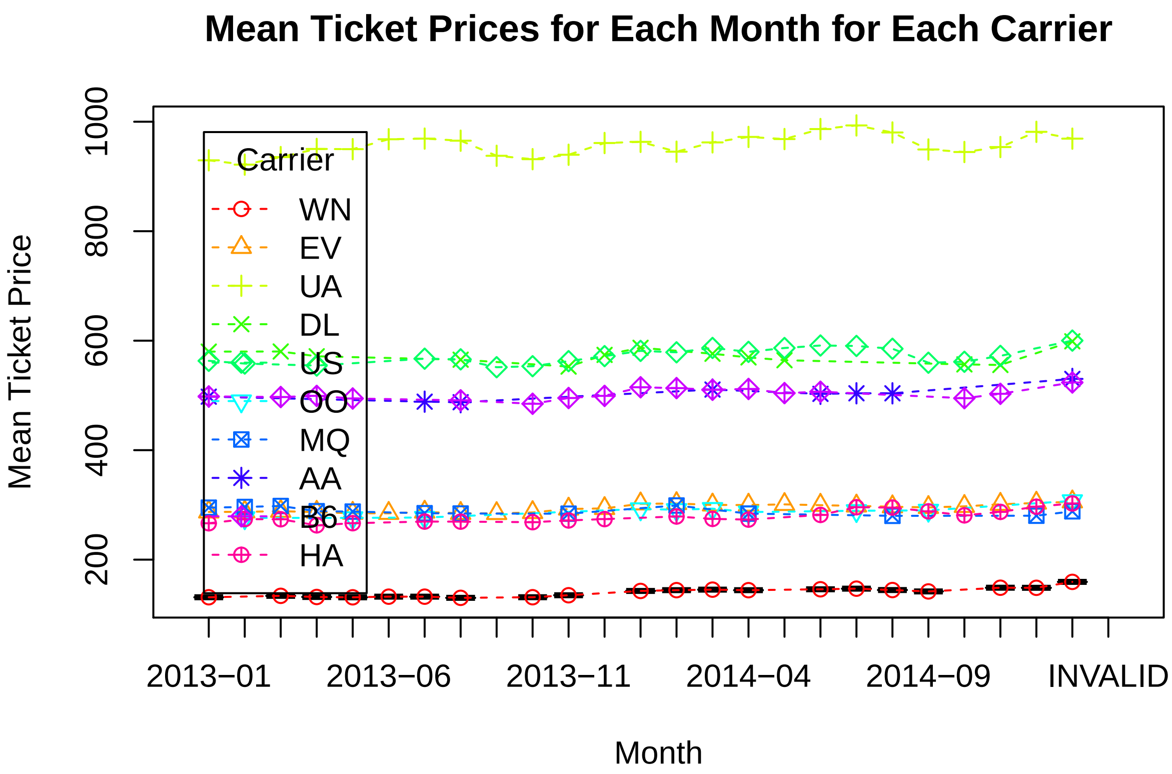
The key value pairs for the same carrier will be combined together before the reduce step. The reducer is responsible for several tasks. For invalid records, the reducer will calculate the total number of the them and the output will be "INVALID" string and number of invalid records pair. For each carrier, the reducer will first go through all the records for it and see if it is still active in 2015. If it's not an active carrier, all the records will be ignored. For active carriers, the number of records will be calculated, then the records will be grouped by month. Mean price for each month is calculated. There will be several outputs for every active carrier, each output is for one month. The key of the output is a string containing month, identifier of the carrier and the total number of flights for this carrier from all given data separated by commas. The value of the output is the mean price for that month for that carrier. Separator between key and value is set to comma so that the output file of the Mapreduce will be in csv format which can be easily read by R for further processing.

Results

The output files from Mapreduce is processed by a R script. Total numbers of flights of all carriers will be sorted. Identifiers of 10 carriers with most flights will be stored in a vector. Then the script uses each identifier in the vector as a filter to find all the outputs for that carrier. These outputs are sorted by month then be ploted on a graph (as shown below).

From the results, we can see that from Jan 2013 to Jan 2015, the monthly average ticket prices didn't change a lot for all the carriers. Though there seems to be a trend that the ticket prices going up slowly through this period of time.

Following is the result graph created by R script.


Mean Ticket Prices for Each Month for Each Carrier
Mean Ticket Price
1000
800
600
400
200
Month
2013−01
2013−06
2013−11
2014−04
2014−09
INVALID
Carrier
WN
EV
UA
DL
US
OO
MQ
AA
B6
HA