Stat250 HW-1

Yu Pei

January 23, 2014

1 0 Stating the Problem

We have 81 csv files of airlines data from 1987 to 2012 september.but the files after 2007 have different format, also the total rows for 81 files are 140 million. So we need an efficient way to compute certain statistics of the Arrival delays columns of the data.

2 Introduction

In this report, I will describe how to use the ArrDelays packages in mygithub account . Which includes two methods to compute mean,median,sd of ArrDelays in all 81 files. And in the SQL section I will describe how to set up database,create table and copy the data into database.Get statistics by linking to postgres from R with RPostgreSQL package.

The methods used in here all exploit the fact that arrival delays are integer numbers thus we can just go over each file once and build up frequency table. Which is efficient regards to both RAM and CPU.

3 Install ArrDelays Package And Usage

To run the function:

1. install the R package, in shell move to the directory with ArrDelays Folder, issue
   R CMD INSTALL -l your/library/address ArrDelays

2. IN R library(ArrDelays) there are 6 functions in this package. main_funcR and main_funcC return the Frequency talbe of class IntegerFrequencyTable. Then you can compute the mean,median,sd, length with the according function defined for IntegerFrequencyTable. The result is what we would want for this question.

4 R-method

The method used in main_funcR is

1. List all the *.csv files in the folder
2. Read one line to check the column of ArrDelay
3. cut the arrdelay column in shell and pipe the result into R with scan function.Then use table function to sort into table.
4. finally merge all the tables as output.

5 C-method

This use a similar method but do it in C

1. Pass the file names to C,(also we can pass in to column numbers to deal with two different format)
2. In C define Struct Table, process each line of the files and put the arrdelay time in the right slot in a Table instance.
3. With R interface .Call, Copy the the table into an R vector and return.

This Method doesn't need to load in the whole file, which is the most time consuming part of the first method.

6 SQL-method

This part we use postgres to construct a database with a table call **delays** with only one column. Loading in data is done with a shell script in my github account. The Loading takes more than 20mins.But the calculation only takes seconds. So if we want to reuse the data, we should use SQL method.

1. The setup of database can see the Reference
2. Create table in postgres **psql** shell use:

   CREATE TABLE delays (arrdelay double precision);
3. Load data with script setup_database.sh. Run it in bash shell with:

   bash setup_database.sh
4. Then with RPostgreSQL package loaded in R with can issue the command:

   freq_table=dbGetQuery(con,'SELECT arrdelay,count(*) FROM delays GROUP BY arrdelay;')