Fial Exam Take Home
CSCE 587
Fall 2016
Due: 12/8/2016 via Dropbox
Name: Sendurr Selvaraj
VIP ID: 00323540

Problem 1:

Expanding on your Hadoop homework, find the mean departure delayby airline. You output should consist of the top 20 mean departure delays by airline in descending order. Assuming the data frame you create is called airlineDelay.df, define the column names to be output using:
colnames(airlineDelay.df) = c('Carrier', 'Delay')

R-code

```
#**********************************
#*  Problem 1                    *
#**********************************
# Set environmental variables
Sys.setenv(HADOOP_CMD="/usr/bin/hadoop")
Sys.setenv(HADOOP_STREAMING="/usr/hdp/2.3.0.0-2557/hadoop-mapreduce/hadoop-streaming-2.7.1.2.3.0.0-2557.jar")

# Load the following packages in the following order
library(rhdfs)
library(rmr2)

# initialize the connection from rstudio to hadoop
hdfs.init()

# Doing simple mapreduce on airline data
# Our map function which returns the keyval < airline, departure delay>
map1 = function(k,flights) {
  return ( keyval(as.character(flights[[9]]),as.numeric(flights[[16]])))
}

# Our reduce function which mean departure delay for each airline
reduce1 = function(airline, delay) {
  keyval(airline, mean(delay,na.rm=TRUE))
}

# Our mapreduce function which invokes map1 and reduce1 and parses
# the input file expected it to be comma delimited
```

```
mr1 = function(input, output = NULL) {
  mapreduce(input = input,
           output = output,
           input.format = make.input.format("csv", sep=","),
           map = map1,
           reduce = reduce1)}

# Set up the input definition (small dataset) and output definition
hdfs.root = '/user/share/student'
hdfs.data = file.path(hdfs.root,'wholeEnchilada.csv')
hdfs.out = file.path(hdfs.root,'out1')

# Invoke out mapreduce job
out = mr1(hdfs.data, hdfs.out)

# Fetch the results from HDFS and coerce into a dataframe
results = from.dfs(out)
results.df = as.data.frame(results, stringsAsFactors=F)

# add column heading to dataframe
colnames(results.df) = c('Carrier', 'Delay')

# Display results
x=results.df[order(-results.df$Delay),]
x[1:20,]
```

Output

```
Carrier   Delay
EV        14.373166
YV        12.918553
B6        11.772661
AA        11.343577
UA        11.194775
FL        10.712942
MQ        10.326245
WN        10.291157
CO        10.008638
AS        9.919690
DH        9.762928
OH        9.310795
XE        9.149135
9E        8.466088
US        8.267891
DL        7.948352
OO        7.452644
```

```
HP      7.348048
9       7.340478
NW      6.814502
```

Problem 2:

Find the meandeparture delay by airline/airport combination. Imagine that we hypothesize that some airline/airport combinations have larger departure delays because of their geographical locations and bad management.

Market:We define an airline/airport combination as a pair of airlinecombined with an airport. We will use the hyphen (`-`) as the separating character when pasting the two market strings together. For example, flights on United Airline (UA) departing Columbia (CAE) would be, UA-CAE. This problem will require you to paste the airline string and origin airport strings together to effectively create a multi-valueairline/airport key.

Hint:the key returned by your map function should be the airline/airport combo. You caneither deal with NA values in
your map function or your reduce function. Assuming the data frame you create is called departureDelay.df, define the column names to be output using: colnames(departureDelay.df) = c('Carrier/Airport', 'Delay')

a)Display your results for the 20 airline/airports having the largest mean departure delays in descending order.

R-code:

```
#***********************************
#*  Problem 2                     *
#***********************************

map2 = function(k,flights) {
#  return ( keyval(as.character(flights[[9]]),flights[[16]]))
  pair=paste(as.character(flights[[9]]), as.character(flights[[17]]), sep="-")
  return ( keyval(pair,as.numeric(flights[[16]])))
}

# Our reduce function which finds the largest taxin time for each destination airports
```

```
reduce2 = function(car_airport, delay) {
  keyval(car_airport, mean(delay,na.rm=TRUE))
}

# Our mapreduce function which invokes map1 and reduce1 and parses
# the input file expected it to be comma delimited
mr2 = function(input, output = NULL) {
  mapreduce(input = input,
            output = output,
            input.format = make.input.format("csv", sep=","),
            map = map2,
            reduce = reduce2)}

# Set up the input definition (small dataset) and output definition
hdfs.root = '/user/share/student'
hdfs.data = file.path(hdfs.root,'wholeEnchilada.csv')
hdfs.out = file.path(hdfs.root,'out2')

# Invoke out mapreduce job
out = mr2(hdfs.data, hdfs.out)

# Fetch the results from HDFS and coerce into a dataframe
results = from.dfs(out)
results.df = as.data.frame(results, stringsAsFactors=F)

# add column heading to dataframe
colnames(results.df) = c('Carrier/Airport', 'Delay')

# Display results
x=results.df[order(-results.df$Delay),]
x[1:20,]
```

Output

```
Carrier/Airport   Delay
OO-SHV          251.60000
OO-FMN          240.00000
```

B6-LAX	224.00000
OO-OGD	172.40000
OO-CYS	105.00000
OO-PUB	104.00000
XE-TWF	100.00000
OH-MCN	59.00000
DH-MSY	58.14286
9-PIR	56.50000
DL-HLN	54.50000
9E-MSO	52.50000
OH-RNO	52.16667
HA-PIT	52.00000
OH-GNV	52.00000
9E-BZN	47.83333
9E-EWR	47.25000
B6-ACK	45.76987
9-MKE	43.55556
AA-SHV	42.00000

b)Does there appear to be pattern of airline/airports having the largest mean departure delays? Based on your results in a), are there any airline/airport combinations that you would want to avoid?

Answer: The largest airline/airports combination consists primarily of regional airports. We can omit airline/airports combination with airlines as 'OO'. 'OO' airlines maximum delay in 5 airports which means the delay is not due to airport location but poor management by the airlines. Therefore there seems to be a problem with the airline and not the location.