EE 5243

Hadoop Project

Submitted By

S M Azharul Karim

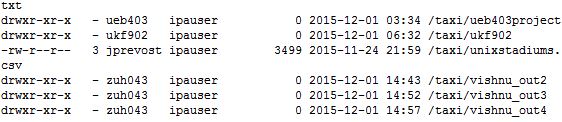
Banner ID- 01519097

**Hadoop Project:**

I've written two python scripts- mapper.py and reducer.py. The mapper file takes user input of some data and maps/sorts it. The reducer takes the output of the mapper and converts the data into a smaller set of data.

In this project, the data is the list of stadiums and their details. A .csv file was provided which contained the stadium data- name of the stadiums, capacity of each stadium turf type and so on. The task is to find how many stadiums have artificial turf. The mapper takes the .csv file as input and sorts the data based on the variable "turf". Here, turf can have two possible values-TRUE, or FALSE. Based on these two values the sorting is done. The reducer then counts how many stadiums have the turf value TRUE assigned to it. That is the number of stadiums that have artificial turf. Finally, the reducer prints its output in an output folder. All the codes are tested and run in the hadoop cluster.

I accessed the /taxi folder and saw the files in there by using the command: hdfs dfs -ls /taxi



**Python codes:**

mapper.py:

#!/usr/bin/env python

import sys

## This code is written by S M Azharul Karim, Banner ID-01519097 ##

## This is the mapper ##

count=0

count1=0

for line in sys.stdin:

line=line.strip()

stadium, capacity, expanded, location, surface, turf, team, opened, weather, roof, elevation = line.split(",")

# writing the output to STDOUT

#output here will be input of reducer

print '%s\t%s' %(turf,1)

reducer.py

#!/usr/bin/env python

from operator import itemgetter

import sys

## this code is written by S M Azharul Karim, Banner ID-01519097 ##

## This is the reducer ##

current\_countart = 0

current\_countnatural = 0

current\_turf=None

turf=None

## input comes from STDIN

for line in sys.stdin:

## remove leading and trailing whitespace

line = line.strip()

## parse the input we got from mapper.py

turf, count= line.split("\t",1)

## convert count (currently a string) to int

try:

count = int(count)

except ValueError:

## count was not a number, so silently

## ignore/discard this line

continue

if current\_turf==turf:

current\_countart=current\_countart+1

## this IF-switch only works because Hadoop sorts map output

## by key (here: word) before it is passed to the reducer

else:

if current\_turf:

current\_countnatural += current\_countnatural

#write result to STDOUT

print '%s\t%s' % (current\_turf, current\_countart)

current\_countart = count

current\_turf = turf

## printing output to output directory

if current\_turf == turf:

print '%s\t%s' % (current\_turf, current\_countart)

print "total no of artificial turfs:",current\_countart

Command to test the mapper and reducer:

cat unixstadiums.csv | ./mapper.py | sort -k1,1 | ./reducer.py

Command for running the mapper-reducer in hadoop cluster:

hadoop jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming-2.7.1.2.3.2.0-2950.jar -file /home/dhf227/mapper.py -mapper /home/dhf227/mapper.py -file /home/dhf227/reducer.py -reducer /home/dhf227/reducer.py -input /taxi/unixstadiums.csv -output /user/dhf227/outfile

Command for getting the output file:

hadoop fs -get outfile1/\* ~/myfiles/

Output file:

FALSE 15

TRUE 17

total no of artificial turfs: 17

(Note: the data file I have in the hadoop has 32 lines. Hence, the total no of output is 15+17=32.)

Screenshots of successful completion of the code in hadoop cluster:



