Hw1 record

Login:

ssh [jl10919@dumbo.hpc.nyu.edu](mailto:jl10919@dumbo.hpc.nyu.edu)

getdata from HDFS:

hfs -get /user/hc2660/hw1data

copy all files under directory:

scp -r ​jl10919​@dumbo.hpc.nyu.edu:~​jl10919​/task1/.

scp ​[jl10919​@dumbo.hpc.nyu.edu:~​jl10919​/task1/TripFareJoin.txt](mailto:jl10919​@dumbo.hpc.nyu.edu:~​jl10919​/task1/TripFareJoin.txt) .

scp ​jl10919​@dumbo.hpc.nyu.edu:~​jl10919​/jl10919\_hw1.zip .

upload file to dumbo (put py code in /src)

scp t1\_mapper.py [jl10919@dumbo.hpc.nyu.edu:~jl10919/task1](mailto:jl10919@dumbo.hpc.nyu.edu:~jl10919/task1)

scp t2\_a\_mapper.py jl10919@dumbo.hpc.nyu.edu:~jl10919/example/src

修改权限，否则permission denied

chmod a+x t1\_mapper.py

upload input file to Hadoop

hfs -put ./trips\_samp.csv

hfs -put ./fares\_samp.csv

check items in Hadoop

hfs -ls

clear output in Hadoop

hfs -rm -r e2.out

run

hjs -D mapreduce.job.reduces=2 -file ~/example/src -mapper src/t1\_mapper.py -reducer src/t1\_reducer.py -input /user/jl10919/trip\_data.csv -input /user/jl10919/fare\_data.csv -output /user/jl10919/t1.out

hjs -D mapreduce.job.reduces=1 -file ~/example/src -mapper src/t1\_mapper.py -reducer src/t1\_reducer.py -input /user/jl10919/trip\_data.csv -input /user/jl10919/fare\_data.csv -output /user/jl10919/t1.out

get output to local vm

hfs -get e2.out

merge output

hfs -getmerge e2.out e2.txt

看前几行

head e1.txt

测试mapper.py

[jl10919@login-1-1 src]$ echo "AA70234AB7643A84903E4B0705352D8A,2A3BEA5321E55025D86D65269A67DDD7,VTS,2013-08-05 23:07:00,CRD,52.0000000000,0.0000000000,0.5000000000,10.4000000000,0.0000000000,62.9000000000" | ./t1\_mapper.py

AA70234AB7643A84903E4B0705352D8A,2A3BEA5321E55025D86D65269A67DDD7,VTS,2013-08-05 23:07:00#CRD,52.0000000000,0.0000000000,0.5000000000,10.4000000000,0.0000000000,62.9000000000,1

Reducer.py test

echo -e "BE63DAC380EDE4C0D59593C0891B569C,101888DE3F330CEFA6B67C29407E95B3,VTS,2013-08-07 08:40:00#1,,2013-08-07 08:51:00,2,660,1.19000,-73.9807360000,40.7479740000,-73.9825670000,40.7602540000,2\nBE63DAC380EDE4C0D59593C0891B569C,101888DE3F330CEFA6B67C29407E95B3,VTS,2013-08-07 08:40:00#CRD,25.0000000000,0.0000000000,0.5000000000,4.0000000000,0.0000000000,29.5000000000,1" | ./t1\_reducer.py

<https://blog.csdn.net/wangato/article/details/70173682>

变量名 & input

<https://blog.csdn.net/modiziri/article/details/47661091>

test mapper

hjs -D mapreduce.job.reduces=2 -file ~/example/src -mapper src/t1\_mapper.py -input /user/jl10919/trips\_samp.csv -input /user/jl10919/fares\_samp.csv -output /user/jl10919/m4.out

t1

hjs -D mapreduce.job.reduces=1 -file ~/example/src -mapper src/t1\_mapper.py -reducer src/t1\_reducer.py -input /user/jl10919/trip\_data.csv -input /user/jl10919/fare\_data.csv -output /user/jl10919/t1.out

hfs -get t1.out TripFareJoin

hfs -getmerge t1.out TripFareJoin.txt

t2 -a

test mapper.py

echo -e "BE63DAC380EDE4C0D59593C0891B569C,101888DE3F330CEFA6B67C29407E95B3,VTS,2013-08-07 08:40:00\t1,,2013-08-07 08:51:00,2,660,1.19000,-73.9807360000,40.7479740000,-73.9825670000,40.7602540000,CRD,25.0000000000,0.0000000000,0.5000000000,4.0000000000,0.0000000000,29.5000000000" | ./t2\_a\_mapper.py

test reducer.py

echo -e "24.01,28\t1\n24.01,28\t1\n8.01,12\t1\n4.01,8\t1" | ./t2\_a\_reducer.py

hjs -D mapreduce.job.reduces=1 -file ~/example/src -mapper src/t2\_a\_mapper.py -reducer src/t2\_a\_reducer.py -input /user/jl10919/TripFareJoin.txt -output /user/jl10919/t2\_a.out

hfs -get t2\_a.out FareAmounts

hjs -D mapreduce.job.reduces=1 -file ~/task2-e -mapper task2-e/map.py -reducer task2-e/reduce.py -input /user/jl10919/TripFareJoin.txt -output FareAmounts

t2 -b

hjs -D mapreduce.job.reduces=1 -file ~/example/src -mapper src/t2\_b\_mapper.py -reducer src/t2\_b\_reducer.py -input /user/jl10919/TripFareJoin.txt -output /user/jl10919/t2\_b.out

hfs -get t2\_b.out TripAmount

t2-c

hjs -D mapreduce.job.reduces=1 -file ~/example/src -mapper src/t2\_c\_mapper.py -reducer src/t2\_c\_reducer.py -input /user/jl10919/TripFareJoin.txt -output /user/jl10919/t2\_c.out

hfs -get t2\_c.out NumberPassengers

t2-d

echo -e " 2013-08-07\t29.000000,0.000000\n2013-08-07\t29.000000,0.000000" | ./t2\_d\_reducer.py

hjs -D mapreduce.job.reduces=1 -file ~/example/src -mapper src/t2\_d\_mapper.py -reducer src/t2\_d\_reducer.py -input /user/jl10919/TripFareJoin.txt -output /user/jl10919/t2\_d.out

hfs -get t2\_d.out TotalRevenue

t2-e

hjs -D mapreduce.job.reduces=1 -file ~/example/src -mapper src/t2\_e\_mapper.py -reducer src/t2\_e\_reducer.py -input /user/jl10919/TripFareJoin.txt -output /user/jl10919/t2\_e.out

hfs -get t2\_e.out MedallionTrips

t2-f

hjs -D mapreduce.job.reduces=1 -file ~/example/src -mapper src/t2\_f\_mapper.py -reducer src/t2\_f\_reducer.py -input /user/jl10919/TripFareJoin.txt -output /user/jl10919/t2\_f.out

hfs -get t2\_f.out UniqueTaxis

t3

hjs -D mapreduce.job.reduces=1 -file ~/example/src -mapper src/t3\_mapper.py -reducer src/t3\_reducer.py -input /user/jl10919/TripFareJoin.txt -input /user/jl10919/licenses.csv -output /user/jl10919/t3.out

hfs -get t3.out VehicleJoin

test mapper.py

echo -e "BE63DAC380EDE4C0D59593C0891B569C,101888DE3F330CEFA6B67C29407E95B3,VTS,2013-08-07 08:40:00\t1,,2013-08-07 08:51:00,2,660,1.19000,-73.9807360000,40.7479740000,-73.9825670000,40.7602540000,CRD,25.0000000000,0.0000000000,0.5000000000,4.0000000000,0.0000000000,29.5000000000" | ./t2\_a\_mapper.py

zip -r jl.zip jl/\*