gedit titanicmapper.py

#!/usr/bin/python

import sys

def mapper():

for line in sys.stdin:

PassengerId,survived,pclass,Name,Sex,Age,SibSp,Parch,Ticket,Fare,Cabin,Embarked = line.strip().split()

if survived == '1':

print 'average\_age\t{0}\t{1}'.format(Sex,Age)

elif survived == '0':

print 'survived\_class\t{0}\t1'.format(pclass)

mapper()

gedit titanic.txt

PassengerId survived pclass Name Sex Age SibSp Parch Ticket Fare Cabin Embarked 1 1 3 "Hockley" male 22 1 0 1100 7.25 c55 S

2 1 1 "Margaret" female 38 1 0 599 7.2833 C85 C

3 0 3 "Laina" female 26 0 0 1282 7.925 c44 S

4 0 2 "Jack" male 20 0 0 1000 7.00 c10 C

5 1 1 "Rose" female 19 1 1 1500 8.25 c11 S

cat titanic.txt | python titanicmapper.py

gedit titanicreducer.py

#!/usr/bin/python TITANIC

import sys

def reducer():

total\_age\_male = 0

total\_age\_female = 0

count\_male = 0

count\_female = 0

survived\_class\_count = {}

for line in sys.stdin:

key, v1,v2 = line.strip().split('\t')

if key == 'average\_age':

sex, age = v1,v2

if sex == 'male':

total\_age\_male += float(age)

count\_male += 1

elif sex=='female':

total\_age\_female += float(age)

count\_female += 1

elif key == 'survived\_class':

pclass, count = v1,v2

if pclass in survived\_class\_count:

survived\_class\_count[pclass] += int(count)

else:

survived\_class\_count[pclass] = int(count)

average\_age\_male = total\_age\_male / count\_male if count\_male > 0 else 0 average\_age\_female = total\_age\_female / count\_female if count\_female > 0 else 0 print'Average Age of Males Who Died:' ,average\_age\_male

print'Average Age of Females Who Died: ',average\_age\_female

for pclass, count in survived\_class\_count.items():

print'Number of Persons Survived in Class ',pclass,':', count

reducer()

cat titanic.txt | python titanicmapper.py | sort | python titanicreducer.py

hdfs dfs -put titanic.txt titanic.txt