499100

Homework 4

1.

Mapper

```
[baoxin.l@ip-172-31-95-86 ~]$ hdfs dfs -put used_cars.csv
[baoxin.l@ip-172-31-95-86 ~]$ hdfs dfs -ls
Found 2 items
drwxrwxrwx   - baoxin.l baoxin.l          0 2021-11-12 22:55 .scratchdir
-rw-r--r--   3 baoxin.l baoxin.l   69145610 2021-12-03 00:36 used_cars.csv
[baoxin.l@ip-172-31-95-86 ~]$ nano aa_mapper.py
[baoxin.l@ip-172-31-95-86 ~]$ cat aa_mapper.py
#!usr/bin/env python
import sys
for line in sys.stdin:
        line = line.strip()
        name = line.split(",")[14]
        price = line.split(",")[18]
        print '%s\t%s' % (name, price)
```

Reduce

```
[baoxin.l@ip-172-31-95-86 ~]$ nano aa_reduce.py
[baoxin.l@ip-172-31-95-86 ~]$
[baoxin.l@ip-172-31-95-86 ~]$ cat aa_reduce.py
import sys

D = {}
Dcount = {}
Dmax = {}
for line in sys.stdin:
        name,price = line.strip().split('\t')
        try:
                price = float(price)
        except ValueError:
                continue
        try:
                D[name] = D[name] + price
        except:
                D[name] = price
        try:
                Dcount[name] = Dcount[name] + 1
        except:
                Dcount[name] = 1
        try:
                Dmax[name]=max(Dmax[name], price)
        except:
                Dmax[name]= price
for maker in D.keys():
        print '%s\t%s\t%s' %(maker, float(D[maker]/Dcount[maker]), float(Dmax[maker]))
```

```
[baoxin.l@ip-172-31-95-86 ~]$ chmod +x aa_mapper.py
[baoxin.l@ip-172-31-95-86 ~]$ chmod +x aa_reduce.py
```

```
[baoxin.l@ip-172-31-95-86 ~]$ cat used_cars.csv |python aa_mapper.py | python aa_reduce.py
Mercury 3824.92857143   5999.0
RAM     39612.2912088   70910.0
Maserati        85833.08        124721.0
MINI    15995.1923077   29550.0
Chevrolet       22369.2841001   115000.0
Porsche 76359.137931    329500.0
Lamborghini     213033.285714   289500.0
Mercedes-Benz   33328.8900145   859000.0
Volkswagen      23553.7661871   51769.0
Saab    5775.4  9990.0
Rolls-Royce     216060.214286   449995.0
Alfa Romeo      44807.4700855   97579.0
Scion   5796.5  11999.0
Cadillac        42457.2492212   99860.0
Honda   16887.214831    89995.0
Hyundai 22205.7184537   49585.0
Ford    29194.807856    79924.0
Mazda   22642.8757515   53095.0
GMC     25946.9153846   78500.0
Spyker  305500.0        305500.0
BMW     43258.8887334   175045.0
Kia     20497.9208473   49330.0
Mitsubishi      13649.5633803   34995.0
smart   11664.75        13888.0
Land Rover      51574.2322581   135465.0
Suzuki  5197.4  7499.0
Dodge   28131.1145663   149000.0
Lincoln 38103.0106383   93190.0
Acura   20452.1824324   44888.0
Jaguar  31863.59375     65375.0
Jeep    29960.4518674   88500.0
Aston Martin    89668.0 89668.0
Nissan  16420.6660175   52332.0
Toyota  23582.4692192   87038.0
Volvo   45222.545657    82375.0
Fisker  33800.0 33800.0
INFINITI        31769.0316206   73930.0
Chrysler        22941.4642857   50001.0
FIAT    12426.2857143   21116.0
McLaren 122221.0        139500.0
Pontiac 6552.1  29995.0
Saturn  3936.66666667   5950.0
Bentley 125571.25       335995.0
Hummer  12637.6666667   16995.0
```

Bash

```
[baoxin.l@ip-172-31-95-86 ~]$ cat aa_bash.sh
hadoop jar /opt/cloudera/parcels/CDH-7.1.7-1.cdh7.1.7.p0.15945976/jars/hadoop-streaming-3.1.1.7.1.7.0-551.ja
r \
        -Dmapred.reduce.tasks=1 \
        -input /user/baoxin.l/used_cars.csv \
        -output /user/baoxin.l/used_cars_output \
        -file aa_mapper.py \
        -file aa_reduce.py \
        -mapper "python aa_mapper.py" \
        -reducer "python aa_reduce.py"
[baoxin.l@ip-172-31-95-86 ~]$
```

```
[baoxin.l@ip-172-31-95-86 ~]$ bash aa_bash.sh
WARNING: Use "yarn jar" to launch YARN applications.
21/12/03 01:47:27 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files ins
tead.
packageJobJar: [aa_mapper.py, aa_reduce.py] [/opt/cloudera/parcels/CDH-7.1.7-1.cdh7.1.7.p0.15945976/jars/had
oop-streaming-3.1.1.7.1.7.0-551.jar] /tmp/streamjob7666496075942172491.jar tmpDir=null
21/12/03 01:47:28 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-95-86.ec2.internal/172.31.
95.86:8032
21/12/03 01:47:28 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-95-86.ec2.internal/172.31.
95.86:8032
21/12/03 01:47:29 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /user/baoxin.l/.sta
ging/job_1636491610632_2738
21/12/03 01:47:29 INFO mapred.FileInputFormat: Total input files to process : 1
21/12/03 01:47:29 INFO mapreduce.JobSubmitter: number of splits:2
21/12/03 01:47:29 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapreduce.
job.reduces
21/12/03 01:47:29 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1636491610632_2738
21/12/03 01:47:29 INFO mapreduce.JobSubmitter: Executing with tokens: []
21/12/03 01:47:29 INFO conf.Configuration: resource-types.xml not found
21/12/03 01:47:29 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
21/12/03 01:47:29 INFO impl.YarnClientImpl: Submitted application application_1636491610632_2738
21/12/03 01:47:29 INFO mapreduce.Job: The url to track the job: http://ip-172-31-95-86.ec2.internal:8088/pro
xy/application_1636491610632_2738/
21/12/03 01:47:29 INFO mapreduce.Job: Running job: job_1636491610632_2738
21/12/03 02:19:19 INFO mapreduce.Job: Job job_1636491610632_2738 running in uber mode : false
21/12/03 02:19:19 INFO mapreduce.Job:  map 0% reduce 0%
21/12/03 02:19:25 INFO mapreduce.Job:  map 100% reduce 0%
21/12/03 02:19:32 INFO mapreduce.Job:  map 100% reduce 100%
21/12/03 02:19:32 INFO mapreduce.Job: Job job_1636491610632_2738 completed successfully
21/12/03 02:19:32 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=99815
                FILE: Number of bytes written=952181
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=69176053
                HDFS: Number of bytes written=1458
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=3
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
```

```
		Launched map tasks=2
		Launched reduce tasks=1
		Data-local map tasks=2
		Total time spent by all maps in occupied slots (ms)=6942
		Total time spent by all reduces in occupied slots (ms)=3219
		Total time spent by all map tasks (ms)=6942
		Total time spent by all reduce tasks (ms)=3219
		Total vcore-milliseconds taken by all map tasks=6942
		Total vcore-milliseconds taken by all reduce tasks=3219
		Total megabyte-milliseconds taken by all map tasks=7108608
		Total megabyte-milliseconds taken by all reduce tasks=3296256
	Map-Reduce Framework
		Map input records=20000
		Map output records=20000
		Map output bytes=251675
		Map output materialized bytes=102094
		Input split bytes=240
		Combine input records=0
		Combine output records=0
		Reduce input groups=53
		Reduce shuffle bytes=102094
		Reduce input records=20000
		Reduce output records=52
		Spilled Records=40000
		Shuffled Maps =2
		Failed Shuffles=0
		Merged Map outputs=2
		GC time elapsed (ms)=201
		CPU time spent (ms)=6140
		Physical memory (bytes) snapshot=1633300480
		Virtual memory (bytes) snapshot=8622391296
		Total committed heap usage (bytes)=2005401600
		Peak Map Physical memory (bytes)=632705024
		Peak Map Virtual memory (bytes)=2870611968
		Peak Reduce Physical memory (bytes)=370388992
		Peak Reduce Virtual memory (bytes)=2881757184
	Shuffle Errors
		BAD_ID=0
		CONNECTION=0
		IO_ERROR=0
		WRONG_LENGTH=0
		WRONG_MAP=0
		WRONG_REDUCE=0
	File Input Format Counters
		Bytes Read=69175813
	File Output Format Counters
		Bytes Written=1458
12/03 02:19:32 INFO streaming.StreamJob: Output directory: /user/baoxin.l/used_cars_output
```

```
[baoxin.l@ip-172-31-95-86 ~]$ hdfs dfs -cat used_cars_output/part-00000|head -10
Mercury 3824.92857143   5999.0
McLaren 122221.0        139500.0
Maserati        85833.08        124721.0
MINI    15995.1923077   29550.0
Chevrolet       22369.2841001   115000.0
Porsche 76359.137931    329500.0
Lamborghini     213033.285714   289500.0
Mercedes-Benz   33328.8900145   859000.0
Volkswagen      23553.7661871   51769.0
Genesis 42795.147541    75695.0
```

2.

Mapper

```
[baoxin.l@ip-172-31-95-86 ~]$ nano accident.py
[baoxin.l@ip-172-31-95-86 ~]$ chmod +x accident.py
[baoxin.l@ip-172-31-95-86 ~]$ cat accident.py
import sys
for line in sys.stdin:
        line = line.strip()
        maker = line.split(',')[14]
        price = line.split(',')[18]
        accident = line.split(',')[9]
        if accident =='FALSE':
                maker = maker +' No Accidents'
                print '%s\t%s' %(maker,price)
        elif accident =='TRUE':
                maker = maker +'With Acciednts'
                print '%s\t%s' %(maker,price)
        else :
                continue
```

```
[baoxin.l@ip-172-31-95-86 ~]$ cat used_cars.csv |python accident.py |python aa_reduce.py
Cadillac No Accidents   29109.2347826   79439.0
NissanWith Acciednts    12437.5228216   26995.0
Rolls-Royce No Accidents        227279.0        449995.0
Jaguar No Accidents     33413.36        65375.0
Chevrolet No Accidents  21545.8843627   115000.0
MercuryWith Acciednts   3954.125        5999.0
ToyotaWith Acciednts    15153.1070111   39998.0
Alfa Romeo No Accidents 34461.0909091   97579.0
DodgeWith Acciednts     17590.0241935   49995.0
Honda No Accidents      15627.2167969   89995.0
BuickWith Acciednts     13494.25        21495.0
smart No Accidents      10923.6666667   11888.0
MINI No Accidents       16702.2352941   29550.0
Mitsubishi No Accidents 12711.0625      34995.0
Lincoln No Accidents    29240.9558824   85995.0
Buick No Accidents      19912.5632184   51275.0
PontiacWith Acciednts   3832.33333333   4999.0
HyundaiWith Acciednts   12464.1377246   28740.0
Lamborghini No Accidents        213033.285714   289500.0
Maserati No Accidents   39376.8857143   75000.0
MazdaWith Acciednts     15718.0588235   32991.0
GenesisWith Acciednts   25333.75        27851.0
Mercedes-Benz No Accidents      34656.9278132   859000.0
SuzukiWith Acciednts    5999.0  5999.0
GMCWith Acciednts       19307.8181818   54900.0
Hummer No Accidents     13185.2 16995.0
JeepWith Acciednts      19401.1473214   79991.0
smartWith Acciednts     13888.0 13888.0
GMC No Accidents        25967.7720207   78500.0
Jeep No Accidents       24616.6217949   88500.0
Lexus No Accidents      28891.5414013   66900.0
Suzuki No Accidents     4997.0  7499.0
Volkswagen No Accidents 16498.2110092   38995.0
FIATWith Acciednts      18999.0 18999.0
Aston Martin No Accidents       89668.0 89668.0
RAM No Accidents        31629.5839416   62995.0
Mazda No Accidents      18668.7131148   32999.0
Pontiac No Accidents    7717.71428571   29995.0
MaseratiWith Acciednts  38811.6666667   47995.0
SaabWith Acciednts      2500.0  2500.0
Volvo No Accidents      29761.1968504   82375.0
Subaru No Accidents     17185.2808511   46995.0
FordWith Acciednts      15207.1535581   63980.0
Mercedes-BenzWith Acciednts     27497.9569892   85800.0
HummerWith Acciednts    9900.0  9900.0
ScionWith Acciednts     5231.0  9995.0
INFINITIWith Acciednts  22936.9811321   41959.0
Audi No Accidents       28704.4736842   159999.0
BentleyWith Acciednts   130650.0        131800.0
HondaWith Acciednts     13525.2896825   37495.0
Spyker No Accidents     305500.0        305500.0
Scion No Accidents      6927.5  11999.0
Nissan No Accidents     15954.0 47995.0
Tesla No Accidents      55328.3333333   82995.0
ChryslerWith Acciednts  11472.1666667   28991.0
FerrariWith Acciednts   139995.0        139995.0
Land RoverWith Acciednts        34901.0952381   79995.0
```

```
Land RoverWith Acciednts        34901.0952381   79995.0
VolkswagenWith Acciednts        12913.8734177   27781.0
Saturn No Accidents     4360.75 5950.0
Chrysler No Accidents   17685.9312977   32500.0
INFINITI No Accidents   25189.4166667   62995.0
Bentley No Accidents    125006.944444   335995.0
Mercury No Accidents    3652.66666667   5499.0
Saab No Accidents       6594.25 9990.0
FIAT No Accidents       11330.8333333   21116.0
Toyota No Accidents     20464.4852507   61899.0
Land Rover No Accidents 42758.1625      97500.0
Porsche No Accidents    84128.1805556   329500.0
TeslaWith Acciednts     55995.0 55995.0
KiaWith Acciednts       12210.2619048   25780.0
JaguarWith Acciednts    26328.7142857   36495.0
RAMWith Acciednts       27779.15625     44995.0
Ford No Accidents       26071.1172566   78900.0
VolvoWith Acciednts     17902.2727273   39425.0
McLaren No Accidents    122221.0        139500.0
AcuraWith Acciednts     17236.5454545   31950.0
SaturnWith Acciednts    3597.4  4995.0
Genesis No Accidents    34773.8695652   45000.0
BMW No Accidents        34685.3306773   102415.0
BMWWith Acciednts       26361.3349057   74800.0
Lotus No Accidents      35900.0 35900.0
CadillacWith Acciednts  22376.4210526   63995.0
Kia No Accidents        15795.2394737   45736.0
Dodge No Accidents      24989.2964072   149000.0
Acura No Accidents      21242.4747475   44888.0
FiskerWith Acciednts    33800.0 33800.0
Hyundai No Accidents    15341.7193347   49087.0
Ferrari No Accidents    533171.333333   1390000.0
Rolls-RoyceWith Acciednts       148747.5        155995.0
PorscheWith Acciednts   39067.7333333   74995.0
MitsubishiWith Acciednts        5619.33333333   14995.0
SubaruWith Acciednts    13645.712766    30995.0
AudiWith Acciednts      24856.6216216   64495.0
Alfa RomeoWith Acciednts        31161.6666667   31495.0
ChevroletWith Acciednts 14492.0147059   51495.0
MINIWith Acciednts      14659.6666667   21995.0
LincolnWith Acciednts   21516.275       36995.0
LexusWith Acciednts     20903.26        39998.0
```

bash

```
[baoxin.l@ip-172-31-95-86 ~]$ nano bb_bash.sh
[baoxin.l@ip-172-31-95-86 ~]$ cat bb_bash.sh
hadoop jar /opt/cloudera/parcels/CDH-7.1.7-1.cdh7.1.7.p0.15945976/jars/hadoop-streaming-3.1.1.7.1.7.0-551.jar \
        -Dmapred.reduce.tasks=1 \
        -input /user/baoxin.l/used_cars.csv \
        -output /user/baoxin.l/accident_output \
        -file accident.py \
        -file aa_reduce.py \
        -mapper "python accident.py" \
        -reducer "python aa_reduce.py"
```

```
[baoxin.l@ip-172-31-95-86 ~]$ bash bb_bash.sh
WARNING: Use "yarn jar" to launch YARN applications.
21/12/03 02:38:51 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [accident.py, aa_reduce.py] [/opt/cloudera/parcels/CDH-7.1.7-1.cdh7.1.7.p0.15945976/jars/hadoop-streaming-3.1.1.7.1.7.0-5
51.jar] /tmp/streamjob6523991732317878 4.jar tmpDir=null
21/12/03 02:38:52 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-95-86.ec2.internal/172.31.95.86:8032
21/12/03 02:38:52 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-95-86.ec2.internal/172.31.95.86:8032
21/12/03 02:38:52 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /user/baoxin.l/.staging/job_1636491610632_2786
21/12/03 02:38:53 INFO mapred.FileInputFormat: Total input files to process : 1
21/12/03 02:38:53 INFO mapreduce.JobSubmitter: number of splits:2
21/12/03 02:38:53 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
21/12/03 02:38:53 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1636491610632_2786
21/12/03 02:38:53 INFO mapreduce.JobSubmitter: Executing with tokens: []
21/12/03 02:38:53 INFO conf.Configuration: resource-types.xml not found
21/12/03 02:38:53 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
21/12/03 02:38:53 INFO impl.YarnClientImpl: Submitted application application_1636491610632_2786
21/12/03 02:38:53 INFO mapreduce.Job: The url to track the job: http://ip-172-31-95-86.ec2.internal:8088/proxy/application_1636491610632
_2786/
21/12/03 02:38:53 INFO mapreduce.Job: Running job: job_1636491610632_2786
21/12/03 02:41:31 INFO mapreduce.Job: Job job_1636491610632_2786 running in uber mode : false
21/12/03 02:41:31 INFO mapreduce.Job:  map 0% reduce 0%
21/12/03 02:41:37 INFO mapreduce.Job:  map 100% reduce 0%
21/12/03 02:41:42 INFO mapreduce.Job:  map 100% reduce 100%
21/12/03 02:41:42 INFO mapreduce.Job: Job job_1636491610632_2786 completed successfully
21/12/03 02:41:42 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=74752
                FILE: Number of bytes written=902870
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=69176053
                HDFS: Number of bytes written=4028
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=6292
                Total time spent by all reduces in occupied slots (ms)=2651
                Total time spent by all map tasks (ms)=6292
                Total time spent by all reduce tasks (ms)=2651
                Total vcore-milliseconds taken by all map tasks=6292
                Total vcore-milliseconds taken by all reduce tasks=2651
                Total megabyte-milliseconds taken by all map tasks=6443008
                Total megabyte-milliseconds taken by all reduce tasks=2714624
        Map-Reduce Framework
                Map input records=20000
                Map output records=12954
                Map output bytes=335539
                Map output materialized bytes=77867
                Input split bytes=240
                Combine input records=0
                Combine output records=0
                Reduce input groups=98
                Reduce shuffle bytes=77867
                Reduce input records=12954
                Reduce output records=98
                Spilled Records=25908
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=149
                CPU time spent (ms)=5450
                Physical memory (bytes) snapshot=1645838336
                Virtual memory (bytes) snapshot=8627875840
                Total committed heap usage (bytes)=2009595904
                Peak Map Physical memory (bytes)=635650048
                Peak Map Virtual memory (bytes)=2872098816
                Peak Reduce Physical memory (bytes)=381181952
                Peak Reduce Virtual memory (bytes)=2884182016
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=69175813
        File Output Format Counters
                Bytes Written=4028
```

```
21/12/03 02:41:42 INFO streaming.StreamJob: Output directory: /user/baoxin.l/accident_output
[baoxin.l@ip-172-31-95-86 ~]$ hdfs dfs -cat accident_output/part-00000|head -20
Chevrolet No Accidents  21545.8843627   115000.0
NissanWith Acciednts    12437.5228216   26995.0
Rolls-Royce No Accidents        227279.0        449995.0
Jaguar No Accidents     33413.36        65375.0
Cadillac No Accidents   29109.2347826   79439.0
MercuryWith Acciednts   3954.125        5999.0
ToyotaWith Acciednts    15153.1070111   39998.0
Alfa Romeo No Accidents 34461.0909091   97579.0
DodgeWith Acciednts     17590.0241935   49995.0
Honda No Accidents      15627.2167969   89995.0
BuickWith Acciednts     13494.25        21495.0
smart No Accidents      10923.6666667   11888.0
MINI No Accidents       16702.2352941   29550.0
Mitsubishi No Accidents 12711.0625      34995.0
Lincoln No Accidents    29240.9558824   85995.0
Buick No Accidents      19912.5632184   51275.0
PontiacWith Acciednts   3832.33333333   4999.0
HyundaiWith Acciednts   12464.1377246   28740.0
Lamborghini No Accidents        213033.285714   289500.0
Maserati No Accidents   39376.8857143   75000.0
[baoxin.l@ip-172-31-95-86 ~]$
```