Assignment 1

Tianzhixi Yin
Big Data

February 22, 2014

Map and Reduce routines

Mapper

I used Pidm as the key, all the other columns as value.

```
#!/usr/bin/python

import sys

for line in sys.stdin:
    data = line.strip().split(”,”)
    Pidm=data[0]
    All=data[1:]
    print ”{0}\t{1}”.format(Pidm, ’,’.join(All))
```

Reducers

My first reducer is for putting all the duplicate lines for one Pidm into one line.

```
#!/usr/bin/python

import sys

oldKey=None
pr=[]

for line in sys.stdin:
    data_mapped = line.strip().split(”\t”)
    if len(data_mapped) != 2:
        # Something has gone wrong. Skip this line.
        continue

    thisKey, All = data_mapped

    if oldKey and oldKey != thisKey:
        print oldKey, ”,”, ',␣'.join(pr)
        oldKey = thisKey;
        pr = []

    oldKey = thisKey
    pr.append(All)

if oldKey != None:
    print oldKey, ”,”, ',␣'.join(pr)
```

My second reducer is for combining the Grade file and the FinAid file using the Pidms. I did not put the Grade data and FinAid data into one line, because I think the number of columns would be too many.

```
#!/usr/bin/python

import sys

for line in sys.stdin:
    data_mapped = line.strip().split(”\t”)
    if len(data_mapped) != 2:
        # Something has gone wrong. Skip this line.
        continue

    Pidm, All = data_mapped
    print Pidm, ”,”, All
```

My last reducer is for finding out the unique Pidms in the Grade file.

```
#!/usr/bin/python

import sys

oKey=None
ooKey=None
oAll=None

for line in sys.stdin:
    data_mapped = line.strip().split(”\t”)
    if len(data_mapped) != 2:
        # Something has gone wrong. Skip this line.
        continue

    thisKey, All = data_mapped
    if oKey!=ooKey and oKey!=thisKey:
        print oKey, ',', oAll

    ooKey = oKey
    oKey=thisKey
    oAll=All
```

Map-Reduce system

I used Hadoop on CentOS 6.3. Hong Zhang helped me to install a single-node cluster. Then I used Hadoop Streaming to run map-reduce job with the Python codes I wrote. The Hadoop version is 2.0.0, the Java version is 1.6.0_31. First I installed Java and set PATH in /ect/profile. Then I installed Hadoop and set PATH in /ect/profile. I set IP for namenode and datanode. I configured core-site.xml, mapred-site.xml and hdfs-site.xml. I created a tmp directory under /usr/hadoop, changed its right to 755. Lastly, I formatted namenode.

Unique Pidms

I found 9770 unique Pidms for the Grade file. I found 0 unique Pidm for the FinAid file. I found 44221 common Pidms for both files.