R08921A16 莊士賢 HW3

I use dataproc on google cloud platform to build a hadoop cluster

use wget to put access log into hdfs

```
karta2108003@cluster-hadoop-m:/usr/lib/hadoop-mapreduce$ sudo wget http://hpc.ee.ntu.edu.tw/html/IntelligentClouds/webAccessLog/access_log
--2020-11-03 08:05:24--  http://hpc.ee.ntu.edu.tw/html/IntelligentClouds/webAccessLog/access_log
Resolving hpc.ee.ntu.edu.tw (hpc.ee.ntu.edu.tw)... 140.112.42.50
Connecting to hpc.ee.ntu.edu.tw (hpc.ee.ntu.edu.tw)|140.112.42.50|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 174449 (170K)
Saving to: 'access_log'

access_log                  100%[=======================================================>] 170.36K  --.-KB/s    in 0.02s

2020-11-03 08:05:25 (8.77 MB/s) - 'access_log' saved [174449/174449]
karta2108003@cluster-hadoop-m:/usr/lib/hadoop-mapreduce$ hdfs dfs -put access_log /tests/input
karta2108003@cluster-hadoop-m:/usr/lib/hadoop-mapreduce$ hdfs dfs -ls /tests/input
Found 1 items
-rw-r--r--   2 karta2108003 hadoop     174449 2020-11-03 08:10 /tests/input/access_log
```

mapper.py

```
#!/usr/bin/env python

import sys

for line in sys.stdin:

    line = line.strip()


    words = line.split()

    for word in words:
        if word[0]=='[' :
            date = word[1:12].split('/')
            hr = word[12:].split(':')[1]
            if date[1] == "Mar":
                date[1] = "03"
            msg = '%s-%s-%s T %s:00:00.000' % \
            (date[2], date[1], date[0], hr)

            print '%s\t%s' % (msg, 1)
```

reducer.py

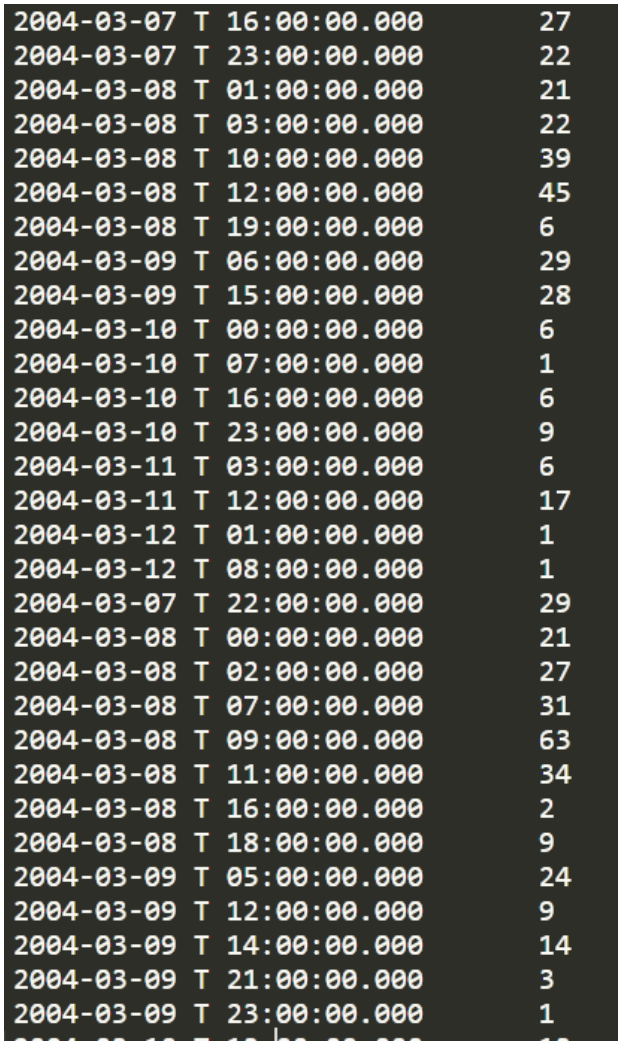
```python
#!/usr/bin/env python

from operator import itemgetter
import sys
current_word = None # 為當前單詞
current_count = 0 # 當前單詞頻數
word = None
for line in sys.stdin:
    words = line.strip() # 去除字串首尾的空白字元
    word, count = words.split('\t') # 按照製表符分隔單詞和數量
    try:
        count = int(count) # 將字串型別的'1'轉換為整型1
    except ValueError:
        continue
    if current_word == word: # 如果當前的單詞等於讀入的單詞
        current_count += count # 單詞頻數加1
    else:
        if current_word: # 如果當前的單詞不為空則列印其單詞和頻數
            print '%s\t%s' %(current_word, current_count)
        current_count = count # 否則將讀入的單詞賦值給當前單詞，且更新頻數
        current_word = word
if current_word == word:
    print '%s\t%s' %(current_word, current_count)
```

start mapreduce

```
karta2108003@cluster-hadoop-m:/usr/lib/hadoop-mapreduce$ hdfs dfs -rmdir /tests/output
karta2108003@cluster-hadoop-m:/usr/lib/hadoop-mapreduce$ hadoop jar hadoop-streaming-2.9.2.jar -D mapreduce.job.reduces=1 -files mapper.py,r
educer.py -mapper mapper.py -reducer reducer.py -input /tests/input/access_log -output /tests/output
packageJobJar: [] [/usr/lib/hadoop-mapreduce/hadoop-streaming-2.9.2.jar] /tmp/streamjob706181822716353876.jar tmpDir=null
20/11/04 08:44:12 INFO client.RMProxy: Connecting to ResourceManager at cluster-hadoop-m/10.140.0.6:8032
20/11/04 08:44:12 INFO client.AHSProxy: Connecting to Application History server at cluster-hadoop-m/10.140.0.6:10200
20/11/04 08:44:13 INFO client.RMProxy: Connecting to ResourceManager at cluster-hadoop-m/10.140.0.6:8032
20/11/04 08:44:13 INFO client.AHSProxy: Connecting to Application History server at cluster-hadoop-m/10.140.0.6:10200
20/11/04 08:44:13 INFO mapred.FileInputFormat: Total input files to process : 1
20/11/04 08:44:13 INFO mapreduce.JobSubmitter: number of splits:15
20/11/04 08:44:13 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.sys
tem-metrics-publisher.enabled
20/11/04 08:44:14 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1604389514237_0003
20/11/04 08:44:14 INFO impl.YarnClientImpl: Submitted application application_1604389514237_0003
20/11/04 08:44:14 INFO mapreduce.Job: The url to track the job: http://cluster-hadoop-m:8088/proxy/application_1604389514237_0003/
20/11/04 08:44:14 INFO mapreduce.Job: Running job: job_1604389514237_0003
20/11/04 08:44:21 INFO mapreduce.Job: Job job_1604389514237_0003 running in uber mode : false
20/11/04 08:44:21 INFO mapreduce.Job:  map 0% reduce 0%
20/11/04 08:44:30 INFO mapreduce.Job:  map 13% reduce 0%
20/11/04 08:44:33 INFO mapreduce.Job:  map 33% reduce 0%
```

result



```
2004-03-07 T 16:00:00.000       27
2004-03-07 T 23:00:00.000       22
2004-03-08 T 01:00:00.000       21
2004-03-08 T 03:00:00.000       22
2004-03-08 T 10:00:00.000       39
2004-03-08 T 12:00:00.000       45
2004-03-08 T 19:00:00.000       6
2004-03-09 T 06:00:00.000       29
2004-03-09 T 15:00:00.000       28
2004-03-10 T 00:00:00.000       6
2004-03-10 T 07:00:00.000       1
2004-03-10 T 16:00:00.000       6
2004-03-10 T 23:00:00.000       9
2004-03-11 T 03:00:00.000       6
2004-03-11 T 12:00:00.000       17
2004-03-12 T 01:00:00.000       1
2004-03-12 T 08:00:00.000       1
2004-03-07 T 22:00:00.000       29
2004-03-08 T 00:00:00.000       21
2004-03-08 T 02:00:00.000       27
2004-03-08 T 07:00:00.000       31
2004-03-08 T 09:00:00.000       63
2004-03-08 T 11:00:00.000       34
2004-03-08 T 16:00:00.000       2
2004-03-08 T 18:00:00.000       9
2004-03-09 T 05:00:00.000       24
2004-03-09 T 12:00:00.000       9
2004-03-09 T 14:00:00.000       14
2004-03-09 T 21:00:00.000       3
2004-03-09 T 23:00:00.000       1
```

github url : https://github.com/destiny3952/cloudcomputinghw3.git