Student: Thuong Han, Truong
ID: 610088

Student: Hoang Thang, Mai
ID: 610089

PROJECT – PART 1-5

Part 1: Set up a single node cluster and optionally an eclipse development environment to create and test your programs

a) Setup environment on Virtual box, Docker and Google Cloud
   Ref Single Hadoop with GoogleCloud Instance.docx
b) Get Cloudera
   Ref Single Hadoop with GoogleCloud Instance.docx
c) WordCount – run on google cloud

Prepare data

echo "Hadoop is an elephant" > file0
echo "Hadoop is as yellow as can be" > file1
echo "Oh what a yellow fellow is Hadoop" > file2
hadoop fs -put file* /user/cloudera/wordcount/input

```
[root@quickstart ~]# echo "Hadoop is an elephant" > file0
[root@quickstart ~]# echo "Hadoop is as yellow as can be" > file1
[root@quickstart ~]# echo "Oh what a yellow fellow is Hadoop" > file2
[root@quickstart ~]# hadoop fs -put file* /user/cloudera/wordcount/input
[root@quickstart ~]#
```

Load jar file into cloudera

```
[root@quickstart ~]# curl -o http://cuidot.vn/data/wordcount.jar
curl: no URL specified!
curl: try 'curl --help' or 'curl --manual' for more information
[root@quickstart ~]# curl -O http://cuidot.vn/data/wordcount.jar
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
101  4282  101  4282    0     0   2629      0  0:00:01  0:00:01 --:--:-- 20198
[root@quickstart ~]# ls
file0  file1  file2  hue.json  wordcount.jar
[root@quickstart ~]#
```

Run wordcount

```
[root@quickstart ~]# curl -o http://cuidot.vn/data/wordcount.jar
curl: no URL specified!
curl: try 'curl --help' or 'curl --manual' for more information
[root@quickstart ~]# curl -O http://cuidot.vn/data/wordcount.jar
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
101  4282  101  4282    0     0   2629      0  0:00:01  0:00:01 --:--:-- 20198
[root@quickstart ~]# ls
file0  file1  file2  hue.json  wordcount.jar
[root@quickstart ~]# hadoop jar wordcount.jar WordCount /user/cloudera/wordcount/input /user/cloudera/wordcount/out
put
19/10/30 09:03:30 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
19/10/30 09:03:31 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement t
he Tool interface and execute your application with ToolRunner to remedy this.
19/10/30 09:03:31 INFO input.FileInputFormat: Total input paths to process : 3
19/10/30 09:03:31 INFO mapreduce.JobSubmitter: number of splits:3
19/10/30 09:03:32 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1572425408290_0001
19/10/30 09:03:33 INFO impl.YarnClientImpl: Submitted application application_1572425408290_0001
19/10/30 09:03:33 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1
572425408290_0001/
19/10/30 09:03:33 INFO mapreduce.Job: Running job: job_1572425408290_0001
```

Wordcount result:

```
[root@quickstart ~]# hadoop jar wordcount.jar WordCount /user/cloudera/wordcount/input /user/cloudera/wordcount/out
put
19/10/30 09:03:30 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
19/10/30 09:03:31 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement t
he Tool interface and execute your application with ToolRunner to remedy this.
19/10/30 09:03:31 INFO input.FileInputFormat: Total input paths to process : 3
19/10/30 09:03:31 INFO mapreduce.JobSubmitter: number of splits:3
19/10/30 09:03:32 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1572425408290_0001
19/10/30 09:03:33 INFO impl.YarnClientImpl: Submitted application application_1572425408290_0001
19/10/30 09:03:33 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1
572425408290_0001/
19/10/30 09:03:33 INFO mapreduce.Job: Running job: job_1572425408290_0001
19/10/30 09:03:49 INFO mapreduce.Job: Job job_1572425408290_0001 running in uber mode : false
19/10/30 09:03:49 INFO mapreduce.Job:  map 0% reduce 0%
19/10/30 09:04:11 INFO mapreduce.Job:  map 33% reduce 0%
19/10/30 09:04:12 INFO mapreduce.Job:  map 67% reduce 0%
19/10/30 09:04:13 INFO mapreduce.Job:  map 100% reduce 0%
19/10/30 09:04:31 INFO mapreduce.Job:  map 100% reduce 100%
19/10/30 09:04:32 INFO mapreduce.Job: Job job_1572425408290_0001 completed successfully
19/10/30 09:04:33 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=200
                FILE: Number of bytes written=574817
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=482
                HDFS: Number of bytes written=80
                HDFS: Number of read operations=12
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=1
                Launched map tasks=3
                Launched reduce tasks=1
                Data-local map tasks=3
                Total time spent by all maps in occupied slots (ms)=60450
                Total time spent by all reduces in occupied slots (ms)=16932
                Total time spent by all map tasks (ms)=60450
                Total time spent by all reduce tasks (ms)=16932
                Total vcore-milliseconds taken by all map tasks=60450
                Total vcore-milliseconds taken by all reduce tasks=16932
                Total megabyte-milliseconds taken by all map tasks=61900800
                Total megabyte-milliseconds taken by all reduce tasks=17338368
        Map-Reduce Framework
                Map input records=3
                Map output records=18
                Map output bytes=158
                Map output materialized bytes=212
                Input split bytes=396
                Combine input records=0
                Combine output records=0
                Reduce input groups=12
                Reduce shuffle bytes=212
                Reduce input records=18
                Reduce output records=12
                Spilled Records=36
                Shuffled Maps =3
                Failed Shuffles=0
                Merged Map outputs=3
                GC time elapsed (ms)=806
                CPU time spent (ms)=3340
                Physical memory (bytes) snapshot=828334080
                Virtual memory (bytes) snapshot=5230444544
```

d) InMapperWordCount – run on google cloud

Prepare data (skip if use current data that wordcount used)

```
echo "Hadoop is an elephant" > file0
echo "Hadoop is as yellow as can be" > file1
echo "Oh what a yellow fellow is Hadoop" > file2
hadoop fs -put file* /user/cloudera/wordcount/input
```

Load jar file into cloudera

```
[root@quickstart /]# curl -O http://cuidot.vn/data/hadoop.jar
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
  3 48.6M    3 1834k    0     0   268k      0  0:03:05  0:00:06  0:02:59  321k
```

Run inMapperWordCount

```
[root@quickstart /]# hadoop jar hadoop.jar edu.mum.bigdata.part1.wordcount.inmapper.InMapperWordCount /user/clouder
a/wordcount/input /user/cloudera/wordcount/output
```

InMapperWordCount Result:

```
[root@quickstart /]# hadoop fs -cat /user/cloudera/wordcount/output/part-r-00000
Hadoop  3
Oh      1
a       1
an      1
as      2
be      1
can     1
elephant        1
fellow  1
is      3
what    1
yellow  2
```

e) Average Computation Algorithm – run on google cloud

Load data into cloudera

```
[root@quickstart /]# curl -O http://cuidot.vn/data/access_log
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  168k  100  168k    0     0   142k      0  0:00:01  0:00:01 --:--:--  171k
[root@quickstart /]#
```

Load jar file into cloudera

```
[root@quickstart /]# curl -O http://cuidot.vn/data/hadoop.jar
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
  3 48.6M    3 1834k    0     0   268k      0  0:03:05  0:00:06  0:02:59  321k
```

Organize folder and copy file into hdfs

```
[root@quickstart /]# hadoop fs -mkdir /user/cloudera/averagecomputation/ /user/cloudera/averagecomputation/noinmapper /user/cloudera/averagecomputation/noinmapper/input
[root@quickstart /]# hadoop fs -put access_log /user/cloudera/averagecomputation/noinmapper/input
```

Run AverageComputation algorithm

```
[root@quickstart /]# hadoop jar hadoop.jar edu.mum.bigdata.part1.averagecomputation.noinmapper.AverageComputation /user/clou
dera/averagecomputation/noinmapper/input /user/cloudera/averagecomputation/noinmapper/output
19/11/14 07:12:53 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
19/11/14 07:12:55 INFO input.FileInputFormat: Total input paths to process : 1
19/11/14 07:12:55 INFO mapreduce.JobSubmitter: number of splits:1
19/11/14 07:12:55 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1573713913689_0003
19/11/14 07:12:56 INFO impl.YarnClientImpl: Submitted application application_1573713913689_0003
19/11/14 07:12:56 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1573713913
689_0003/
19/11/14 07:12:56 INFO mapreduce.Job: Running job: job_1573713913689_0003
```

AverageComputation results:

```
lj1212.inktomisearch.com      3169.0
lj1216.inktomisearch.com        209.0
lj1220.inktomisearch.com        209.0
lj1223.inktomisearch.com        1941.5
lj1231.inktomisearch.com        209.0
lordgun.org     2869.0
mail.geovariances.fr    6012.217391304348
market-mail.panduit.com 3427.344827586207
mcl02.cnc.bc.ca 10879.5
mmscrm07-2.sac.overture.com     68.0
mth-fgw.ballarat.edu.au 5448.714285714285
nb-bolz.cremona.polimi.it       2300.0
ns.mou.cz       2300.0
ns.wtbts.org    2311.3333333333335
ns3.vonroll.ch  5971.666666666667
ogw.netinfo.bg  2758.0
osdlab.eic.nctu.edu.tw  269.0
p213.54.168.132.tisdip.tiscali.de       5785.75
p5083cd5d.dip0.t-ipconnect.de   7368.0
pc-030-040.eco.rug.nl   7368.0
pc3-registry-stockholm.telia.net        9452.692307692309
pd95f99f2.dip.t-dialin.net      2869.0
```

```
pd9e50809.dip.t-dialin.net      2869.0
pd9e761cf.dip.t-dialin.net      2300.0
pd9eb1396.dip.t-dialin.net      2300.0
pntn02m05-129.bctel.ca  3095.0
pool-68-160-195-60.ny325.east.verizon.net       3724.2
ppp2.p33.is.com.ua      3582.0
proxy0.haifa.ac.il      3271.157894736842
prxint-sxb2.e-i.net     4022.0
prxint-sxb3.e-i.net     9254.07142857143
px7wh.vc.shawcable.net  7649.0
rouble.cc.strath.ac.uk  2869.0
spica.ukc.ac.uk 1973.5
spot.nnacorp.com        4632.4
trrc02m01-40.bctel.ca   3071.75
ts04-ip92.hevanet.com   4431.5
ts05-ip44.hevanet.com   7854.5625
user-0c8hdkf.cable.mindspring.com       5372.2
vlp181.vlp.fi   2869.0
watchguard.cgmatane.qc.ca       5741.0
wc03.mtnk.rnc.net.cable.rogers.com      10936.0
wc09.mtnk.rnc.net.cable.rogers.com      10860.666666666666
wwwcache.lanl.gov       2869.0
yongsan-cache.korea.army.mil    3056.0
```

f) InMapper Average Computation Algorithm – run on google cloud

Load data into cloudera

```
[root@quickstart /]# curl -O http://cuidot.vn/data/access_log
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  168k  100  168k    0     0   142k      0  0:00:01  0:00:01 --:--:--  171k
[root@quickstart /]#
```

Load jar file into cloudera

```
[root@quickstart /]# curl -O http://cuidot.vn/data/hadoop.jar
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
  3 48.6M    3 1834k    0     0   268k      0  0:03:05  0:00:06  0:02:59  321k
```

Organize folder and copy file into hdfs

```
[root@quickstart /]# hadoop fs -mkdir /user/cloudera/averagecomputation/inmapper /user/cloudera/averagecomputation/inmapper/
input
[root@quickstart /]# hadoop fs -put access_log /user/cloudera/averagecomputation/inmapper/input
```

Run AverageComputation algorithm

```
[root@quickstart /]# hadoop jar hadoop.jar edu.mum.bigdata.part1.averagecomputation.inmapper.InMapperAverageComputation /use
r/cloudera/averagecomputation/inmapper/input /user/cloudera/averagecomputation/inmapper/output
19/11/14 07:23:47 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
19/11/14 07:23:49 INFO input.FileInputFormat: Total input paths to process : 1
19/11/14 07:23:49 INFO mapreduce.JobSubmitter: number of splits:1
19/11/14 07:23:50 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1573713913689_0004
19/11/14 07:23:50 INFO impl.YarnClientImpl: Submitted application application_1573713913689_0004
19/11/14 07:23:50 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1573713913
689_0004/
19/11/14 07:23:50 INFO mapreduce.Job: Running job: job_1573713913689_0004
19/11/14 07:24:00 INFO mapreduce.Job: Job job_1573713913689_0004 running in uber mode : false
19/11/14 07:24:00 INFO mapreduce.Job:  map 0% reduce 0%
19/11/14 07:24:08 INFO mapreduce.Job:  map 100% reduce 0%
19/11/14 07:24:15 INFO mapreduce.Job:  map 100% reduce 100%
19/11/14 07:24:16 INFO mapreduce.Job: Job job_1573713913689_0004 completed successfully
19/11/14 07:24:16 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=5903
                FILE: Number of bytes written=299847
                FILE: Number of read operations=0
```

AverageComputation results:

```
[root@quickstart /]# hadoop fs -cat
/user/cloudera/averagecomputation/inmapper/output/part-r-00000
0x503e4fce.virnxx2.adsl-dhcp.tele.dk    1315.6666666666667
1-320.cnc.bc.ca 10879.5
```

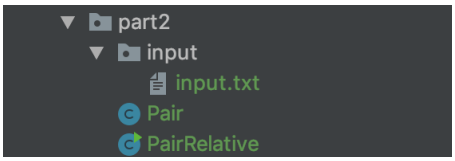
```
1-729.cnc.bc.ca 3262.5714285714284
10.0.0.153      4444.981481481482
12.22.207.235   7368.0
128.227.88.79   5841.785714285715
142.27.64.35    1923.5714285714287
145.253.208.9   3728.285714285714
1513.cps.virtua.com.br  309.0
194.151.73.43   10879.5
195.11.231.210  6032.0
195.230.181.122 2300.0
195.246.13.119  5128.583333333333
2-110.cnc.bc.ca 7912.363636363636
2-238.cnc.bc.ca 3169.0
200-55-104-193.dsl.prima.net.ar 2179.4615384615386
200.160.249.68.bmf.com.br       6634.5
200.222.33.33   2300.0
203.147.138.233 2164.3076923076924
206-15-133-153.dialup.ziplink.net       0.0
206-15-133-154.dialup.ziplink.net       0.0
206-15-133-181.dialup.ziplink.net       0.0
207.195.59.160  4053.05
208-186-146-13.nrp3.brv.mn.frontiernet.net      1689.0
208-38-57-205.ip.cal.radiant.net        3830.3636363636365
208.247.148.12  3067.0
212.21.228.26   2869.0
212.92.37.62    5212.928571428572
213.181.81.4    7649.0
216-160-111-121.tukw.qwest.net  2317.5
216.139.185.45  6051.0
219.95.17.51    3169.0
3_343_lt_someone        6277.2
4.37.97.186     2446.0
61.165.64.6     3056.0
61.9.4.61       2645.3333333333335
64-249-27-114.client.dsl.net    7368.0
```

Student: Thuong Han, Truong
ID: 610088

Student: Hoang Thang, Mai
ID: 610089

Part 2: Implement Pairs algorithm to compute relative frequencies.

There are the .java files for the program.



Run in Hadoop Docker

Show input data content

```
hadoop fs –cat /usr/local/input/part2/input.txt
```

```
bash-4.1# hadoop fs –cat /usr/local/input/part2/input.txt
 B11 C31 D76 A12 B11 C31 D76 C31 A10 B12 D76 C31
D76 D76 B12 B11 C31 D76 B12  C31 B11 A12 C31 B12bash-4.1#
```

Execute the MapReduce program

```
hadoop jar project.jar edu.mum.bigdata.part2.PairRelative
/usr/local/input/part2/input.txt /usr/local/input/part2/output
```

```
bash-4.1# hadoop jar project.jar edu.mum.bigdata.part2.PairRelative /usr/local/input/part2/input.txt /usr/local/input/part2/output
19/11/13 14:39:42 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
19/11/13 14:39:45 INFO input.FileInputFormat: Total input paths to process : 1
19/11/13 14:39:46 INFO mapreduce.JobSubmitter: number of splits:1
19/11/13 14:39:46 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1573529104459_0015
19/11/13 14:39:47 INFO impl.YarnClientImpl: Submitted application application_1573529104459_0015
19/11/13 14:39:48 INFO mapreduce.Job: The url to track the job: http://90b8c019f7a2:8088/proxy/application_1573529104459_0015/
19/11/13 14:39:48 INFO mapreduce.Job: Running job: job_1573529104459_0015
19/11/13 14:40:04 INFO mapreduce.Job: Job job_1573529104459_0015 running in uber mode : false
19/11/13 14:40:04 INFO mapreduce.Job:  map 0% reduce 0%
19/11/13 14:40:18 INFO mapreduce.Job:  map 100% reduce 0%
19/11/13 14:40:33 INFO mapreduce.Job:  map 100% reduce 100%
19/11/13 14:40:34 INFO mapreduce.Job: Job job_1573529104459_0015 completed successfully
19/11/13 14:40:34 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=4494
                FILE: Number of bytes written=239737
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=217
                HDFS: Number of bytes written=643
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=12085
                Total time spent by all reduces in occupied slots (ms)=10779
                Total time spent by all map tasks (ms)=12085
                Total time spent by all reduce tasks (ms)=10779
                Total vcore-seconds taken by all map tasks=12085
                Total vcore-seconds taken by all reduce tasks=10779
                Total megabyte-seconds taken by all map tasks=12375040
                Total megabyte-seconds taken by all reduce tasks=11037696
        Map-Reduce Framework
                Map input records=2
                Map output records=264
                Map output bytes=3960
                Map output materialized bytes=4494
                Input split bytes=121
                Combine input records=0
                Combine output records=0
                Reduce input groups=37
                Reduce shuffle bytes=4494
                Reduce input records=264
                Reduce output records=31
                Spilled Records=528
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=135
                CPU time spent (ms)=3460
                Physical memory (bytes) snapshot=396017664
                Virtual memory (bytes) snapshot=1471606784
                Total committed heap usage (bytes)=227540992
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
```

Program output

```
hadoop fs -cat /usr/local/input/part2/output/*
```

Student: Thuong Han, Truong
ID: 610088

Student: Hoang Thang, Mai
ID: 610089

```
bash-4.1# hadoop fs -cat /usr/local/input/part2/output/*
A10,B12 0.3333333333333333
A10,C31 0.3333333333333333
A10,D76 0.3333333333333333
A12,A10 0.1
A12,B11 0.1
A12,B12 0.2
A12,C31 0.4
A12,D76 0.2
B11,A10 0.06896551724137931
B11,A12 0.10344827586206896
B11,B11 0.06896551724137931
B11,B12 0.1724137931034483
B11,C31 0.3793103448275862
B11,D76 0.20689655172413793
B12,A12 0.125
B12,B11 0.1875
B12,B12 0.1875
B12,C31 0.375
B12,D76 0.125
C31,A10 0.09375
C31,A12 0.09375
C31,B11 0.09375
C31,B12 0.21875
C31,C31 0.28125
C31,D76 0.21875
D76,A10 0.047619047619047616
D76,A12 0.09523809523809523
D76,B11 0.14285714285714285
D76,B12 0.23809523809523808
D76,C31 0.3333333333333333
D76,D76 0.14285714285714285
```

Run in Cloudera environment

```
# ./runAll.sh part2/ edu.mum.bigdata.part2.PairRelative
```

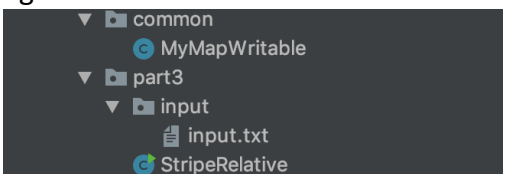
Student: Thuong Han, Truong
ID: 610088

Student: Hoang Thang, Mai
ID: 610089

Part 3: Implement Stripes algorithm to compute relative frequencies.

Run in Hadoop Docker

There are the .java files of program.



Build the whole project into .jar file and copy it into Hadoop docket and execute these commands:

Show the input content

```
hadoop fs –cat /usr/local/input/part3/input/input.txt
```

```
bash-4.1#
bash-4.1# hadoop fs –cat /usr/local/input/part3/input/input.txt

B11 C31 D76 A12 B11 C31 D76 C31 A10 B12 D76 C31
D76 D76 B12 B11 C31 D76 B12  C31 B11 A12 C31 B12
```

Execute the MapReduce program

```
hadoop jar project.jar edu.mum.bigdata.part3.StripeRelative
/usr/local/input/part3/input/input.txt /usr/local/input/part3/output
```

Student: Thuong Han, Truong
ID: 610088

Student: Hoang Thang, Mai
ID: 610089

```
bash-4.1# hadoop jar project.jar edu.mum.bigdata.part3.StripeRelative /usr/local/input/part3/input/input.txt /usr/local/input/part3/output
19/11/12 01:13:14 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
19/11/12 01:13:17 INFO input.FileInputFormat: Total input paths to process : 1
19/11/12 01:13:17 INFO mapreduce.JobSubmitter: number of splits:1
19/11/12 01:13:17 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1573529104459_0004
19/11/12 01:13:17 INFO impl.YarnClientImpl: Submitted application application_1573529104459_0004
19/11/12 01:13:18 INFO mapreduce.Job: The url to track the job: http://90b8c019f7a2:8088/proxy/application_1573529104459_0004/
19/11/12 01:13:18 INFO mapreduce.Job: Running job: job_1573529104459_0004
19/11/12 01:13:28 INFO mapreduce.Job: Job job_1573529104459_0004 running in uber mode : false
19/11/12 01:13:28 INFO mapreduce.Job:  map 0% reduce 0%
19/11/12 01:13:37 INFO mapreduce.Job:  map 100% reduce 0%
19/11/12 01:13:46 INFO mapreduce.Job:  map 100% reduce 100%
19/11/12 01:13:47 INFO mapreduce.Job: Job job_1573529104459_0004 completed successfully
19/11/12 01:13:48 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=2414
                FILE: Number of bytes written=235609
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=226
                HDFS: Number of bytes written=555
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=7041
                Total time spent by all reduces in occupied slots (ms)=6643
                Total time spent by all map tasks (ms)=7041
                Total time spent by all reduce tasks (ms)=6643
                Total vcore-seconds taken by all map tasks=7041
                Total vcore-seconds taken by all reduce tasks=6643
                Total megabyte-seconds taken by all map tasks=7209984
                Total megabyte-seconds taken by all reduce tasks=6802432
        Map-Reduce Framework
                Map input records=2
                Map output records=24
                Map output bytes=2360
                Map output materialized bytes=2414
                Input split bytes=127
                Combine input records=0
                Combine output records=0
                Reduce input groups=6
                Reduce shuffle bytes=2414
                Reduce input records=24
                Reduce output records=6
                Spilled Records=48
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=116
                CPU time spent (ms)=2450
                Physical memory (bytes) snapshot=408604672
                Virtual memory (bytes) snapshot=1478815744
                Total committed heap usage (bytes)=250609664
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
```

```
hadoop fs -cat /usr/local/input/part3/output/*
```

Program output

```
bash-4.1# hadoop fs -cat /usr/local/input/part3/output/*
A10     [B12 0.3333333333333333,D76 0.3333333333333333,C31 0.3333333333333333]
A12     [A10 0.1,B12 0.2,B11 0.1,D76 0.2,C31 0.4]
B11     [A12 0.10344827586206896,A10 0.06896551724137931,B12 0.1724137931034483,B11 0.06896551724137931,D76 0.20689655172413793,C31 0.3793103448275862]
B12     [A12 0.125,B12 0.1875,B11 0.1875,D76 0.125,C31 0.375]
C31     [A10 0.09375,A12 0.09375,B12 0.21875,B11 0.09375,D76 0.21875,C31 0.28125]
D76     [A12 0.09523809523809523,A10 0.047619047619047616,B12 0.23809523809523808,B11 0.14285714285714285,D76 0.14285714285714285,C31 0.3333333333333333]
bash-4.1#
```

Run in Cloudera environment

```
# ./runAll.sh part3/ edu.mum.bigdata.part3.StripeRelative
```

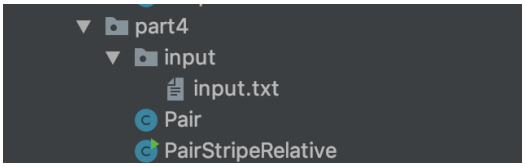
Student: Thuong Han, Truong
ID: 610088

Student: Hoang Thang, Mai
ID: 610089

Part 4: Implement Pairs in Mapper and Stripes in Reducer to compute relative frequencies (Hybrid)

Run in Hadoop Docker

These are the .java files for the program.



Build the whole project into .jar file and copy it to Hadoop docker and execute these commands:

Show the content of input files

```
hadoop fs -cat /usr/local/input/part4/input.txt
```

```
bash-4.1# hadoop fs -cat /usr/local/input/part4/input.txt
B11 C31 D76 A12 B11 C31 D76 C31 A10 B12 D76 C31
D76 D76 B12 B11 C31 D76 B12  C31 B11 A12 C31 B12
```

Execute the MapRecure program

```
hadoop jar project.jar edu.mum.bigdata.part4.PairStripeRelative
/usr/local/input/part4/input.txt /usr/local/input/part4/output
```

Student: Thuong Han, Truong
ID: 610088

Student: Hoang Thang, Mai
ID: 610089

```
bash-4.1# hadoop jar project.jar edu.mum.bigdata.part4.PairStripeRelative /usr/local/input/part4/input.txt /usr/local/input/part4/output
19/11/13 17:45:31 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
19/11/13 17:45:35 INFO input.FileInputFormat: Total input paths to process : 1
19/11/13 17:45:35 INFO mapreduce.JobSubmitter: number of splits:1
19/11/13 17:45:36 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1573529104459_0016
19/11/13 17:45:36 INFO impl.YarnClientImpl: Submitted application application_1573529104459_0016
19/11/13 17:45:37 INFO mapreduce.Job: The url to track the job: http://90b8c019f7a2:8088/proxy/application_1573529104459_0016/
19/11/13 17:45:37 INFO mapreduce.Job: Running job: job_1573529104459_0016
19/11/13 17:45:51 INFO mapreduce.Job: Job job_1573529104459_0016 running in uber mode : false
19/11/13 17:45:51 INFO mapreduce.Job:  map 0% reduce 0%
19/11/13 17:46:01 INFO mapreduce.Job:  map 100% reduce 0%
19/11/13 17:46:15 INFO mapreduce.Job:  map 100% reduce 100%
19/11/13 17:46:17 INFO mapreduce.Job: Job job_1573529104459_0016 completed successfully
19/11/13 17:46:17 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=384
                FILE: Number of bytes written=232229
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=220
                HDFS: Number of bytes written=375
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=6826
                Total time spent by all reduces in occupied slots (ms)=11332
                Total time spent by all map tasks (ms)=6826
                Total time spent by all reduce tasks (ms)=11332
                Total vcore-seconds taken by all map tasks=6826
                Total vcore-seconds taken by all reduce tasks=11332
                Total megabyte-seconds taken by all map tasks=6989824
                Total megabyte-seconds taken by all reduce tasks=11603968
        Map-Reduce Framework
                Map input records=2
                Map output records=27
                Map output bytes=324
                Map output materialized bytes=384
                Input split bytes=121
                Combine input records=0
                Combine output records=0
                Reduce input groups=27
                Reduce shuffle bytes=384
                Reduce input records=27
                Reduce output records=6
                Spilled Records=54
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=125
                CPU time spent (ms)=3400
                Physical memory (bytes) snapshot=418537472
                Virtual memory (bytes) snapshot=1468334080
                Total committed heap usage (bytes)=275775488
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
```

Show the program output

```
hadoop fs -cat /usr/local/input/part4/output/*
```

```
bash-4.1# hadoop fs -cat /usr/local/input/part4/output/*
A10     [B12 0.33333334,D76 0.33333334,C31 0.33333334]
A12     [A10 0.1,B12 0.2,B11 0.1,D76 0.2,C31 0.4]
B11     [A12 0.11764706,A10 0.05882353,B12 0.1764706,D76 0.23529412,C31 0.4117647]
B12     [A12 0.11111111,B11 0.22222222,D76 0.22222222,C31 0.44444445]
C31     [A12 0.16666667,A10 0.083333336,B12 0.25,B11 0.16666667,D76 0.33333334]
D76     [A12 0.125,A10 0.0625,B12 0.25,B11 0.1875,C31 0.375]
```

Run in Cloudera environment

```
# ./runAll.sh part4/ edu.mum.bigdata.part4.PairStripeRelative
```

Part 5: Solve a MapReduce problem of your choice!

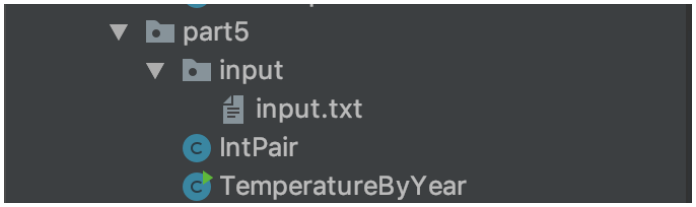
Run in Hadoop Docker

Student: Thuong Han, Truong
ID: 610088

Student: Hoang Thang, Mai
ID: 610089

The problem is to compute the average temperature by year. Build the whole project into .jar file and copy it into Hadoop docker and execute with the follow commands.

The following is the .java files for this part and the testing input data.



Show the content of input data.

```
hadoop fs -cat /usr/local/input/part5/input.txt
```

```
bash-4.1# hadoop fs -cat /usr/local/input/part5/input.txt
1992,12
2008,38
1992,29
2000,29
1992,39
1993,10
1998,15
1992,20
2000,25
1999,14
1997,13
1996,20
1998,23
1999,22
1999,22
2000,21
1993,29
1993,21
2008,25
1992,14
2000,30
1992,30
1997,39
1994,32
```

Execute the MapReduce program

```
hadoop jar project.jar edu.mum.bigdata.part5.TemperatureByYear
/usr/local/input/part5/input.txt /usr/local/input/part5/output
```

Student: Thuong Han, Truong
ID: 610088

Student: Hoang Thang, Mai
ID: 610089

```
bash-4.1# hadoop jar project.jar edu.mum.bigdata.part5.TemperatureByYear /usr/local/input/part5/input.txt /usr/local/input/part5/output
19/11/12 14:06:43 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
19/11/12 14:06:46 INFO input.FileInputFormat: Total input paths to process : 1
19/11/12 14:06:47 INFO mapreduce.JobSubmitter: number of splits:1
19/11/12 14:06:47 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1573529104459_0012
19/11/12 14:06:48 INFO impl.YarnClientImpl: Submitted application application_1573529104459_0012
19/11/12 14:06:48 INFO mapreduce.Job: The url to track the job: http://90b8c019f7a2:8088/proxy/application_1573529104459_0012/
19/11/12 14:06:48 INFO mapreduce.Job: Running job: job_1573529104459_0012
19/11/12 14:07:00 INFO mapreduce.Job: Job job_1573529104459_0012 running in uber mode : false
19/11/12 14:07:00 INFO mapreduce.Job:  map 0% reduce 0%
19/11/12 14:07:09 INFO mapreduce.Job:  map 100% reduce 0%
19/11/12 14:07:18 INFO mapreduce.Job:  map 100% reduce 100%
19/11/12 14:07:19 INFO mapreduce.Job: Job job_1573529104459_0012 completed successfully
19/11/12 14:07:19 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=366
                FILE: Number of bytes written=232163
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=339
                HDFS: Number of bytes written=99
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=6473
                Total time spent by all reduces in occupied slots (ms)=6939
                Total time spent by all map tasks (ms)=6473
                Total time spent by all reduce tasks (ms)=6939
                Total vcore-seconds taken by all map tasks=6473
                Total vcore-seconds taken by all reduce tasks=6939
                Total megabyte-seconds taken by all map tasks=6628352
                Total megabyte-seconds taken by all reduce tasks=7105536
        Map-Reduce Framework
                Map input records=25
                Map output records=24
                Map output bytes=312
                Map output materialized bytes=366
                Input split bytes=121
                Combine input records=0
                Combine output records=0
                Reduce input groups=9
                Reduce shuffle bytes=366
                Reduce input records=24
                Reduce output records=9
                Spilled Records=48
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=70
                CPU time spent (ms)=1890
                Physical memory (bytes) snapshot=419827712
                Virtual memory (bytes) snapshot=1477537792
                Total committed heap usage (bytes)=274202624
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
```

Show the output data.

```
hadoop fs -cat /usr/local/input/part5/output/*
```

```
bash-4.1# hadoop fs -cat /usr/local/input/part5/output/*
1992    24.0
1993    20.0
1994    32.0
1996    20.0
1997    26.0
1998    19.0
1999    19.333334
2000    26.25
2008    31.5
```

Run in Cloudera environment

```
# ./runAll.sh part5/ edu.mum.bigdata.part5.TemperatureByYear
```