Hadoop ile Patent Verisi Islemek

75-99 yillari arasinda hangi patentin hangi hangi patentlere referans verdigi ve patentler hakkinda detayli verileri Hadoop ile isleyecegiz. Veriler alttaki baglantidan alinabilir, gerekli dosyalar Dosyalar cite75_99.txt ve apat63_99.txt

http://www.nber.org/patents/

Referans verisine bakarsak,

```
!head -10 $HOME/Downloads/cite75_99.txt
```

```
"CITING","CITED"
3858241,956203
3858241,1324234
3858241,3398406
3858241,3557384
3858241,3634889
3858242,1515701
3858242,3319261
3858242,3668705
3858242,3707004
```

Detayli patent verisine bakalim

```
!head -10 $HOME/Downloads/apat63_99.txt
```

```
"PATENT","GYEAR","GDATE","APPYEAR","COUNTRY","POSTATE","ASSIGNEE","ASSCODE","CLAIMS","NCLASS","CAT","SUB
3070801,1963,1096,,"BE","",,1,,269,6,69,,1,,0,,,,,,
3070802,1963,1096,,"US","TX",,1,,2,6,63,,0,,,,,,,,
3070803,1963,1096,,"US","IL",,1,,2,6,63,,9,,0.3704,,,,,,
3070804,1963,1096,,"US","OH",,1,,2,6,63,,3,,0.6667,,,,,,
3070805,1963,1096,,"US","CA",,1,,2,6,63,,1,,0,,,,,,
3070806,1963,1096,,"US","PA",,1,,2,6,63,,0,,,,,,,,
3070807,1963,1096,,"US","OH",,1,,623,3,39,,3,,0.4444,,,,,,
3070808,1963,1096,,"US","IA",,1,,623,3,39,,4,,0.375,,,,,,
3070809,1963,1096,,"US","AZ",,1,,4,6,65,,0,,,,,,,,
```

Simdi patent detay verisinden bir orneklem (sample) alalim. Daha ufak bir veri kumesiyle calismak ilk basta faydali olabilir, gelistirme test etme surecini hizlandirir.

```
!chmod a+r $HOME/Downloads/apat63_99.txt
!head -1 $HOME/Downloads/apat63_99.txt > $HOME/Downloads/apat63_99_sampled.txt
!cat $HOME/Downloads/apat63_99.txt | perl -n -e 'print if (rand() < .05)' >> $HOME/
    Downloads/apat63_99_sampled.txt
```

```
Hadoop baslatalim
```

```
!ssh localhost -l hduser /home/hduser/Downloads/hadoop*/bin/stop-all.sh
!ssh localhost -l hduser /home/hduser/Downloads/hadoop*/bin/start-all.sh
```

```
no jobtracker to stop

localhost: no tasktracker to stop

no namenode to stop

localhost: no datanode to stop

localhost: no secondarynamenode to stop

starting namenode, logging to /home/hduser/Downloads/hadoop-1.0.4/libexec/../logs/hadoop-hduser-namenode

localhost: starting datanode, logging to /home/hduser/Downloads/hadoop-1.0.4/libexec/../logs/hadoop-hdus

localhost: starting secondarynamenode, logging to /home/hduser/Downloads/hadoop-1.0.4/libexec/../logs/ha

starting jobtracker, logging to /home/hduser/Downloads/hadoop-1.0.4/libexec/../logs/hadoop-hduser-jobtra

localhost: starting tasktracker, logging to /home/hduser/Downloads/hadoop-1.0.4/libexec/../logs/hadoop-h
```

```
/home/hduser/Downloads/hadoop*/bin/hadoop dfs -mkdir /user/hduser/patent
```

```
!ssh localhost -l hduser /home/hduser/Downloads/hadoop*/bin/hadoop dfs -ls /user/hduser/
    patent
```

```
Found 2 items
-rw-r--r--   1 hduser supergroup  236903179 2013-02-21 14:16 /user/hduser/patent/apat63_99.txt
-rw-r--r--   1 hduser supergroup   11878646 2013-02-21 16:36 /user/hduser/patent/apat63_99_sampled.txt
```

```
!ssh localhost -l hduser /home/hduser/Downloads/hadoop*/bin/hadoop dfs -copyFromLocal /
    home/burak/Downloads/apat63_99_sampled.txt /user/hduser/patent/apat63_99_sampled.txt
```

```
copyFromLocal: Target /user/hduser/patent/apat63_99_sampled.txt already exists
```

Amacimiz patent verisindeki ulke (country) kodunu kullanarak her ulke basina ortalama ne kadar patent uretildigini hesaplamak. Esleme-Indirgeme (Map-Reduce) dongusunde esleme kismini yapacak program asagida.

```
print open("mapper.py").read()
```

```
#!/usr/bin/python
import os,sys
```

```
os.environ['MPLCONFIGDIR']='/tmp'
import pandas as pd
data = pd.read_csv(sys.stdin,sep=",",index_col=0,usecols=[0,4,8])
df = data[pd.notnull(data.ix[:,0]) & pd.notnull(data.ix[:,1])].ix[:,0:2]
df.to_csv(sys.stdout,sep="\t",index=False,header=False)
```

```
!cp mapper.py /tmp/
!chmod a+r /tmp/mapper.py
!chmod a+x /tmp/mapper.py
```

Indirgeyici yazmadan once programimizi iki sekilde test edelim. Bu sekillerden birisi hic indirgeyici olmadan, ikincisi IdentityReducer denen kendisine gecilen veriyi oldugu gibi disari atan (ama yine de ortaa bir indirgeyici oldugu icin sonradan bazi islemlerin yine de yapilacagi) seklinde.

```
!ssh localhost -l hduser /home/hduser/Downloads/hadoop*/bin/hadoop dfs -rmr /user/hduser
    /output
!ssh localhost -l hduser /home/hduser/Downloads/hadoop*/bin/hadoop jar /home/hduser/
    Downloads/hadoop*/contrib/streaming/hadoop-*streaming*.jar -input patent/
    apat63_99_sampled.txt -output output -mapper /tmp/mapper.py -numReduceTasks 0
```

```
Deleted hdfs://localhost:54310/user/hduser/output

packageJobJar: [/app/hadoop/tmp/hadoop-unjar2555196345671652661/] [] /tmp/streamjob5013687273729997973.j

13/02/24 16:30:26 INFO util.NativeCodeLoader: Loaded the native-hadoop library
13/02/24 16:30:26 WARN snappy.LoadSnappy: Snappy native library not loaded
13/02/24 16:30:26 INFO mapred.FileInputFormat: Total input paths to process : 1

13/02/24 16:30:27 INFO streaming.StreamJob: getLocalDirs(): [/app/hadoop/tmp/mapred/local]
13/02/24 16:30:27 INFO streaming.StreamJob: Running job: job_201302241611_0012
13/02/24 16:30:27 INFO streaming.StreamJob: To kill this job, run:
13/02/24 16:30:27 INFO streaming.StreamJob: /home/hduser/Downloads/hadoop-1.0.4/libexec/../bin/hadoop jo
13/02/24 16:30:27 INFO streaming.StreamJob: Tracking URL: http://localhost:50030/jobdetails.jsp?jobid=jo

13/02/24 16:30:28 INFO streaming.StreamJob:  map 0%  reduce 0%

13/02/24 16:30:43 INFO streaming.StreamJob:  map 100%  reduce 0%

13/02/24 16:30:49 INFO streaming.StreamJob:  map 100%  reduce 100%
13/02/24 16:30:49 INFO streaming.StreamJob: Job complete: job_201302241611_0012
13/02/24 16:30:49 INFO streaming.StreamJob: Output: output
```

```
!ssh localhost -l hduser /home/hduser/Downloads/hadoop*/bin/hadoop dfs -copyToLocal
    output /tmp/
```

```
!head -30 /tmp/output/part-00000
```

```
FR 12.0
US 5.0
US 1.0
US 4.0
US 4.0
US 21.0
US 4.0
US 8.0
US 7.0
US 11.0
DE 12.0
US 30.0
US 14.0
US 11.0
US 5.0
JP 21.0
US 23.0
US 5.0
CH 14.0
DE 11.0
US 4.0
US 14.0
US 4.0
US 1.0
US 4.0
IT 3.0
US 1.0
US 7.0
US 8.0
US 6.0
```

Ustteki sonucta goruyoruz ki anahtarlar uretilmis, ama ciktilar anahtara gore siralanmamislar. Hatta ustteki sira girdi sirasiyla tipatip ayni. Simdi <code>IdentityReducer</code> uzerinden.

```
!ssh localhost -l hduser /home/hduser/Downloads/hadoop*/bin/hadoop dfs -rmr /user/hduser
    /output
!ssh localhost -l hduser /home/hduser/Downloads/hadoop*/bin/hadoop jar /home/hduser/
    Downloads/hadoop*/contrib/streaming/hadoop-*streaming*.jar -input patent/
    apat63_99_sampled.txt -output output -mapper /tmp/mapper.py -reducer org.apache.
    hadoop.mapred.lib.IdentityReducer -numReduceTasks 1
```

```
Deleted hdfs://localhost:54310/user/hduser/output

packageJobJar: [/app/hadoop/tmp/hadoop-unjar2314287838929839696/] [] /tmp/streamjob5231815242060775825.j

13/02/24 18:03:14 INFO util.NativeCodeLoader: Loaded the native-hadoop library
13/02/24 18:03:14 WARN snappy.LoadSnappy: Snappy native library not loaded
13/02/24 18:03:14 INFO mapred.FileInputFormat: Total input paths to process : 1
```

```
13/02/24 18:03:14 INFO streaming.StreamJob: getLocalDirs(): [/app/hadoop/tmp/mapred/local]
13/02/24 18:03:14 INFO streaming.StreamJob: Running job: job_201302241759_0004
13/02/24 18:03:14 INFO streaming.StreamJob: To kill this job, run:
13/02/24 18:03:14 INFO streaming.StreamJob: /home/hduser/Downloads/hadoop-1.0.4/libexec/../bin/hadoop jo
13/02/24 18:03:14 INFO streaming.StreamJob: Tracking URL: http://localhost:50030/jobdetails.jsp?jobid=jo

13/02/24 18:03:15 INFO streaming.StreamJob:  map 0%  reduce 0%

13/02/24 18:03:28 INFO streaming.StreamJob:  map 50%  reduce 0%

13/02/24 18:03:31 INFO streaming.StreamJob:  map 100%  reduce 0%

13/02/24 18:03:40 INFO streaming.StreamJob:  map 100%  reduce 100%

13/02/24 18:03:46 INFO streaming.StreamJob: Job complete: job_201302241759_0004
13/02/24 18:03:46 INFO streaming.StreamJob: Output: output
```

```
!ssh localhost -l hduser /home/hduser/Downloads/hadoop*/bin/hadoop dfs -copyToLocal
    output /tmp/
```

```
13/02/24 17:59:51 INFO hdfs.DFSClient: No node available for block: blk_-1575974087486179659_1352 file=/
13/02/24 17:59:51 INFO hdfs.DFSClient: Could not obtain block blk_-1575974087486179659_1352 from any nod
```

```
!head -10 /tmp/output/part-00000
```

```
FR 12.0
US 5.0
US 1.0
US 4.0
US 4.0
US 21.0
US 4.0
US 8.0
US 7.0
US 11.0
```

Ustteki sonucta anahtarlarin siralanmis oldugunu goruyoruz.

```
print open("reducer.py").read()
```

```
#!/usr/bin/python
import os,sys
os.environ['MPLCONFIGDIR']='/tmp'
import pandas as pd
data = pd.read_csv(sys.stdin,sep="\t",names=['country','count'])
grouped = data.groupby('country').mean()
grouped.to_csv(sys.stdout,sep="\t",header=False)
```

```
cat /tmp/output/part-00000 | ./reducer.py | tail -10
```

```
SE 9.0021739130434781
SG 14.0
SU 6.4136125654450264
SV 6.5
TR 8.0
TW 6.2037037037037033
US 10.964136780650541
VE 9.3333333333333339
YU 5.75
ZA 11.170212765957446
```

```
!cp reducer.py /tmp/
!chmod a+r /tmp/reducer.py
!chmod a+x /tmp/reducer.py
```

```
!ssh localhost -l hduser /home/hduser/Downloads/hadoop*/bin/hadoop dfs -rmr /user/hduser
    /output
!ssh localhost -l hduser /home/hduser/Downloads/hadoop*/bin/hadoop jar /home/hduser/
    Downloads/hadoop*/contrib/streaming/hadoop-*streaming*.jar -input patent/
    apat63_99_sampled.txt -output output -mapper /tmp/mapper.py -reducer /tmp/reducer.py
    -numReduceTasks 1
```

```
Deleted hdfs://localhost:54310/user/hduser/output

packageJobJar: [/app/hadoop/tmp/hadoop-unjar3358838375062006941/] [] /tmp/streamjob3628714134396896316.j

13/02/24 20:33:10 INFO util.NativeCodeLoader: Loaded the native-hadoop library
13/02/24 20:33:10 WARN snappy.LoadSnappy: Snappy native library not loaded
13/02/24 20:33:10 INFO mapred.FileInputFormat: Total input paths to process : 1

13/02/24 20:33:10 INFO streaming.StreamJob: getLocalDirs(): [/app/hadoop/tmp/mapred/local]
13/02/24 20:33:10 INFO streaming.StreamJob: Running job: job_201302241759_0005
13/02/24 20:33:10 INFO streaming.StreamJob: To kill this job, run:
13/02/24 20:33:10 INFO streaming.StreamJob: /home/hduser/Downloads/hadoop-1.0.4/libexec/../bin/hadoop jo
13/02/24 20:33:10 INFO streaming.StreamJob: Tracking URL: http://localhost:50030/jobdetails.jsp?jobid=jo

13/02/24 20:33:11 INFO streaming.StreamJob:  map 0%  reduce 0%

13/02/24 20:33:23 INFO streaming.StreamJob:  map 50%  reduce 0%

13/02/24 20:33:26 INFO streaming.StreamJob:  map 100%  reduce 0%

13/02/24 20:33:35 INFO streaming.StreamJob:  map 100%  reduce 100%

13/02/24 20:33:42 INFO streaming.StreamJob: Job complete: job_201302241759_0005
13/02/24 20:33:42 INFO streaming.StreamJob: Output: output
```

```
!ssh localhost -l hduser /home/hduser/Downloads/hadoop*/bin/hadoop dfs -copyToLocal
    output /tmp/
```

```
copyToLocal: Target /tmp/output/_SUCCESS already exists
```

```
!head -10 /tmp/output/part-00001
```

```
JP 8.0
JP 1.0
US 34.0
US 19.0
JP 2.0
US 4.0
US 43.0
GB 6.0
US 14.0
GB 1.0
```