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/

```
!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
```

```
!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
```

```
!chmod a+r $HOME/Downloads/apat63_99.txt
!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,,,,,,,,,
```

```
!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
```

```
print open("pat2.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 pat2.py /tmp/
!chmod a+r /tmp/pat2.py
!chmod a+x /tmp/pat2.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/pat2.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
```

```
!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/pat2.py -reducer org.apache.hadoop.
    mapred.lib.IdentityReducer -numReduceTasks 1
```

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

packageJobJar: [/app/hadoop/tmp/hadoop-unjar4791053218220591275/] [] /tmp/streamjob2130002507404697820.j

13/02/24 16:29:31 INFO util.NativeCodeLoader: Loaded the native-hadoop library
13/02/24 16:29:31 WARN snappy.LoadSnappy: Snappy native library not loaded
13/02/24 16:29:31 INFO mapred.FileInputFormat: Total input paths to process : 1

13/02/24 16:29:31 INFO streaming.StreamJob: getLocalDirs(): [/app/hadoop/tmp/mapred/local]
13/02/24 16:29:31 INFO streaming.StreamJob: Running job: job_201302241611_0011
13/02/24 16:29:31 INFO streaming.StreamJob: To kill this job, run:
13/02/24 16:29:31 INFO streaming.StreamJob: /home/hduser/Downloads/hadoop-1.0.4/libexec/../bin/hadoop jo
13/02/24 16:29:31 INFO streaming.StreamJob: Tracking URL: http://localhost:50030/jobdetails.jsp?jobid=jo

13/02/24 16:29:32 INFO streaming.StreamJob:  map 0%  reduce 0%

13/02/24 16:29:45 INFO streaming.StreamJob:  map 50%  reduce 0%

13/02/24 16:29:48 INFO streaming.StreamJob:  map 100%  reduce 0%

13/02/24 16:29:57 INFO streaming.StreamJob:  map 100%  reduce 100%

13/02/24 16:30:03 INFO streaming.StreamJob: Job complete: job_201302241611_0011
13/02/24 16:30:03 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
```

```
AE 12.0
AG 24.0
AN 15.0
AR 9.0
AR 16.0
AR 2.0
AR 19.0
AR 19.0
AR 11.0
AR 11.0
AR 4.0
AR 10.0
AR 6.0
AR 11.0
AR 8.0
AR 19.0
AR 22.0
AR 1.0
AR 5.0
AR 3.0
AR 10.0
AR 10.0
AR 7.0
AR 11.0
AR 24.0
AR 12.0
AR 3.0
AR 6.0
AT 16.0
AT 7.0
```