Parallel Grep via a Command-Line-Based MapReduce

Jing Xu

jaling@gmail.com

W261-3

1.12.2 Quiz Week 1

1/13/16

In [112]:

```
%%writefile mapper.py
#!/usr/bin/python
import sys
import re
count = 0
WORD_RE = re.compile(r"[\w']+", re.IGNORECASE)
filename = sys.argv[2]
findword = sys.argv[1]
with open (filename, "r") as myfile:
#Please insert your code
    for line in myfile.readlines():
        if findword in line:
            count+=1
print count
```

```
Overwriting mapper.py
```

In [113]:

```
!python mapper.py the LICENSE.txt
```

```
251
```

In [114]:

```
!chmod a+x mapper.py
```

In [115]:

```
%%writefile reducer.py
#!/usr/bin/python
import sys
total = 0
for line in sys.stdin:
    total += int(line)
    #Please insert your code
print total
```

```
Overwriting reducer.py
```

In [116]:

```
!chmod a+x reducer.py
```

In [117]:

```
%%writefile pGrepCount.sh
ORIGINAL_FILE=$1
FIND_WORD=$2
BLOCK_SIZE=$3
CHUNK_FILE_PREFIX=$ORIGINAL_FILE.split
SORTED_CHUNK_FILES=$CHUNK_FILE_PREFIX*.sorted
usage()
{
    echo Parallel grep
    echo usage: pGrepCount filename word chuncksize
    echo greps file file1 in $ORIGINAL_FILE and counts the number o
f lines
    echo Note: file1 will be split in chunks up to $ BLOCK_SIZE chu
nks each
    echo $FIND_WORD each chunk will be grepCounted in parallel
}
#Splitting $ORIGINAL_FILE INTO CHUNKS
split -b $BLOCK_SIZE $ORIGINAL_FILE $CHUNK_FILE_PREFIX
#DISTRIBUTE
for file in $CHUNK_FILE_PREFIX*
do
    #grep -i $FIND_WORD $file|wc -l >$file.intermediateCount &
    ./mapper.py $FIND_WORD $file >$file.intermediateCount &
done
wait
#MERGEING INTERMEDIATE COUNT CAN TAKE THE FIRST COLUMN AND TOTOL...
#numOfInstances=$(cat *.intermediateCount | cut -f 1 | paste -sd+ -
|bc)
numOfInstances=$(cat *.intermediateCount | ./reducer.py)
echo "found [$numOfInstances] [$FIND_WORD] in the file [$ORIGINAL_F
ILE]"
for file in $CHUNK_FILE_PREFIX*
do
    rm "$file"
done
```

```
Overwriting pGrepCount.sh
```

In [118]:

```
!chmod a+x pGrepCount.sh
```

In [119]:

```
!./pGrepCount.sh License.txt COPYRIGHT 4k
```

```
found [11] [COPYRIGHT] in the file [License.txt]
```

In [ ]: