# DATASCI W261: Machine Learning at Scale

David Rose
david.rose@berkeley.edu
W261-1
Week 01
2015.08.31

# This notebook provides a poor man Hadoop through command-line and python. Please insert the python code by yourself.

# Map

In [15]:
```
%%writefile mapper.py
#!/usr/bin/python
import sys
import re
count = 0
WORD_RE = re.compile(r"[\w']+")
filename = sys.argv[2]
findword = sys.argv[1]
wc = 0
with open (filename, "r") as myfile:
    #Please insert your code
    for line in myfile:
        # if line contains the specified word, increment the counte
r
        if re.search(findword, line, re.I) != None:
            wc += 1
print wc
```

```
Overwriting mapper.py
```

In [8]:
```
!chmod a+x mapper.py
```

# Reduce

In [16]:
```
%%writefile reducer.py
#!/usr/bin/python
import sys
sum = 0
for line in sys.stdin:
    #Please insert your code
    # convert the string count to an int and increment the sum
    wc = int(line)
    sum += wc
print sum
```

```
Overwriting reducer.py
```

In [12]:
```
!chmod a+x reducer.py
```

# Write script to file

In [13]:
```
%%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
#MERGING 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]"
```

```
Overwriting pGrepCount.sh
```

# Run the file

In [5]:
```
!chmod a+x pGrepCount.sh
```

Usage: usage: pGrepCount filename word chuncksize

In [17]:
```
!./pGrepCount.sh License.txt COPYRIGHT 4k
```

```
found [57] [COPYRIGHT] in the file [License.txt]
```