DivideAndConquer

September 5, 2016

1 DATASCI W261: Machine Learning at Scale

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

2 Map

```
In [1]: %%writefile mapper.py
        #!/usr/bin/python
        import sys

        findword, filename = sys.argv[1], sys.argv[2]

        with open (filename, "r") as myfile:
            print(sum([1 for line in myfile if findword in line]))
```

```
Overwriting mapper.py
```

```
In [2]: !chmod a+x mapper.py
```

3 Reduce

```
In [3]: %%writefile reducer.py
        #!/usr/bin/python
        import sys

        print(sum([int(line) for line in sys.stdin]))
```

```
Overwriting reducer.py
```

```
In [4]: !chmod a+x reducer.py
```

4 Write script to file

```
In [5]: %%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 chunksize
            echo greps file file1 in $ORIGINAL_FILE and counts the number of lines
            echo Note: file1 will be split in chunks up to $ BLOCK_SIZE chunks 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 COUNTS CAN TAKE THE FIRST COLUMN AND TOTAL...
        #numOfInstances=$(cat *.intermediateCount | cut -f 1 | paste -sd+ - |bc)
        numOfInstances=$(cat *.intermediateCount | ./reducer.py)

        #CLEAN UP
        rm $CHUNK_FILE_PREFIX*

        echo "found [$numOfInstances] [$FIND_WORD] in the file [$ORIGINAL_FILE]"
```

```
Overwriting pGrepCount.sh
```

5 Run the file

```
In [6]: !chmod a+x pGrepCount.sh
```

Usage: usage: pGrepCount filename word chuncksize

```
In [7]: !./pGrepCount.sh License.txt COPYRIGHT 4k
```

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