Prof R. Madana Mohana

BIG DATA ANALYTICS LAB

Exp. Implement following applications using MAPREDUCE on single node cluster

(i) Word Count Application

Input File: word.txt

Bus, Car, bus, car, train, car, bus, car, train, bus,
TRAIN,BUS, buS, caR, CAR, car, BUS, TRAIN

Mapper Code (Python): wcmap.py

```
"""wcmap.py"""
import sys
# input comes from STDIN (standard input)
for line in sys.stdin:
    # remove leading and trailing whitespace
    line = line.strip()
    # split the line into words
    words = line.split()
    # increase counters
    for word in words:
        # write the results to STDOUT (standard output);
        # what we output here will be the input for the
        # Reduce step, i.e. the input for reducer.py
        #
        # tab-delimited; the trivial word count is 1
        print('%s\t%s' %(word, 1))
```

Reducer Code (Python): wcred.py

```
"""wcred.py"""
from operator import itemgetter
import sys

current_word = None
current_count = 0
word = None

# input comes from STDIN
for line in sys.stdin:
    # remove leading and trailing whitespace
    line = line.strip()

    # parse the input we got from mapper.py
    word, count = line.split('\t', 1)

    # convert count (currently a string) to int
    try:
        count = int(count)
    except ValueError:
        # count was not a number, so silently
        # ignore/discard this line
        continue

    # this IF-switch only works because Hadoop sorts map output
    # by key (here: word) before it is passed to the reducer
    if current_word == word:
        current_count += count
    else:
        if current_word:
            # write result to STDOUT
            print('%s\t%s' % (current_word, current_count))
        current_count = count
        current_word = word

# do not forget to output the last word if needed!
if current_word == word:
    print ('%s\t%s' % (current_word, current_count))
```

Step wise Execution:

```
hadoop@ubuntu22:~$ bash
hadoop@ubuntu22:~$ cd Desktop
hadoop@ubuntu22:~/Desktop$ cd MapReduce_wordcount
hadoop@ubuntu22:~/Desktop/MapReduce_wordcount$ ls
wc_python

hadoop@ubuntu22:~/Desktop/MapReduce_wordcount$ cd wc_python
hadoop@ubuntu22:~/Desktop/MapReduce_wordcount/wc_python$ ls
myfile-ip  python-mapreduce-instructions  text  wcmap.py
wcred.py  word.txt

hadoop@ubuntu22:~/Desktop/MapReduce_wordcount/wc_python$ cat
word.txt

Bus, Car, bus,  car, train, car, bus, car, train, bus,
TRAIN,BUS, buS, caR, CAR, car, BUS, TRAIN
```

Running in Local Mode:

```
hadoop@ubuntu22:~/Desktop/MapReduce_wordcount/wc_python$ python3
wcmap.py <word.txt

Bus, 1
Car, 1
bus, 1
car, 1
train,    1
car, 1
bus, 1
car, 1
train,    1
bus, 1
TRAIN,BUS,     1
buS, 1
caR, 1
CAR, 1
car, 1
BUS, 1
TRAIN     1
```

```
hadoop@ubuntu22:~/Desktop/MapReduce_wordcount/wc_python$ cat
word.txt | python3 wcmap.py | sort -k1,1 | python3 wcred.py

bus, 3
buS, 1
Bus, 1
BUS, 1
car, 4
caR, 1
Car, 1
CAR, 1
train,    2
TRAIN     1
TRAIN,BUS,     1
```

Running in Single Node Cluster (Hadoop Environment):

```
hadoop@ubuntu22:~/Desktop/MapReduce_wordcount/wc_python$ start-all.sh

Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [ubuntu22]
Starting resourcemanager
Starting nodemanagers

hadoop@ubuntu22:~/Desktop/MapReduce_wordcount/wc_python$ jps
4566 NameNode
5303 NodeManager
4954 SecondaryNameNode
4731 DataNode
5659 Jps
5164 ResourceManager

hadoop@ubuntu22:~/Desktop/MapReduce_wordcount/wc_python$ hadoop fs -mkdir -p /wordcount

hadoop@ubuntu22:~/Desktop/MapReduce_wordcount/wc_python$ hadoop fs -copyFromLocal word.txt /wordcount

hadoop@ubuntu22:~/Desktop/MapReduce_wordcount/wc_python$ hadoop jar /home/hadoop/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.6.jar -file wcmap.py -mapper "python3 wcmap.py" -file wcred.py -reducer "python3 wcred.py" -input /wordcount/word.txt -output /wordcount/output2

2024-08-12 10:51:12,309 INFO streaming.StreamJob: Output directory: /wordcount/output2
```

OUTPUT:

```
hadoop@ubuntu22:~/Desktop/MapReduce_wordcount/wc_python$ hadoop fs -cat /wordcount/output2/part-00000
BUS, 1
Bus, 1
CAR, 1
Car, 1
TRAIN     1
TRAIN,BUS,     1
buS, 1
bus, 3
caR, 1
car, 4
train,    2
```