Writing an Hadoop MapReduce Program in Python

Shadi Ibrahim
March 30th, 2017

Exercise 1: My Wordcount application in python

Here are example programs in Python which implement wordcount:

```
#!/usr/bin/env python

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)
```

Figure 1: mapper.py

```python
#!/usr/bin/env python

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)
```

Figure 2: reducer.py

Make sure the file has execution permission `chmod +x mapper.py AND chmod +x reducer.py`.

Question 1.1

Generate 2 GB of data and run wordcount application using the follwoing commend:
$ `bin/hadoop jar contrib/streaming/hadoop-*streaming*.jar -mapper /path/mapper.py -reducer /path/reducer.py -input path-to-input -output path-to-output`

Question 1.2

Run wordcount on 2 GB of data while changing the number of reducers and enabling and disabling the combiner.