HADOOP STREAMING

1. Install Python

```
# apt update

# apt install python-is-python3

# whereis python3
```

Hoặc

```
# apt update && sudo apt upgrade -y

# apt install software-properties-common -y

# add-apt-repository ppa:deadsnakes/ppa -y

# add-apt-repository ppa:deadsnakes/nightly -y

# apt update

# apt install python3.11

# python3.11 --version
```

2. Example Using Python WordCount

- **Mapper Phase Code**

  Tạo file mapper.py và cấp quyền chmod +x mapper.py

```
#!/usr/bin/python3
"""mapper.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))
```

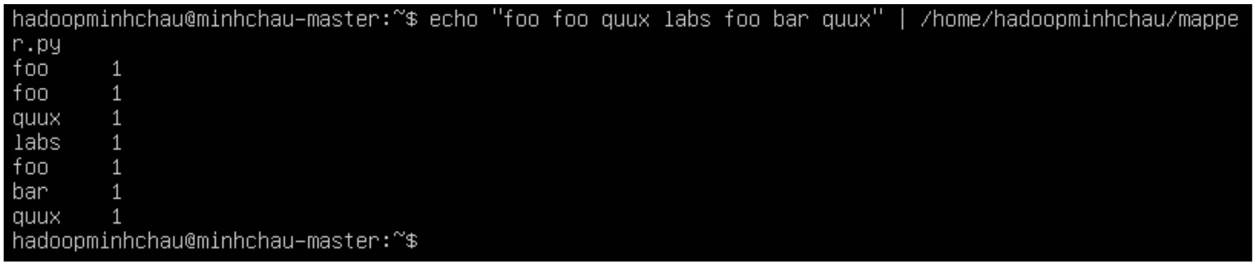
- **Reducer Phase Code**

  Tạo file reducer.py và cấp quyền chmod +x reducer.py

```
#!/usr/bin/python3
"""reducer.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))
```

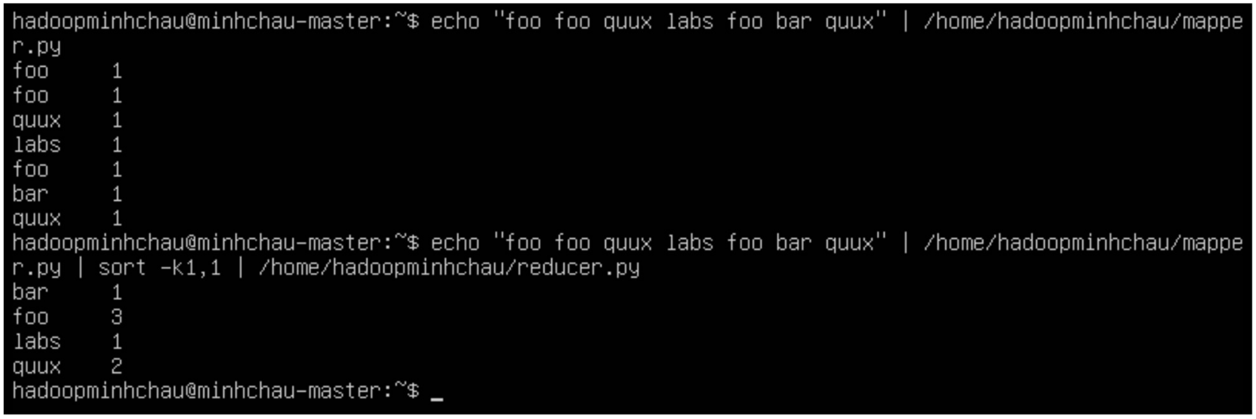
3. Thực thi chương trình WordCount trên thư mục cục bộ

```
$ echo "foo foo quux labs foo bar quux" |
/home/hadoopminhchau/mapper.py
```



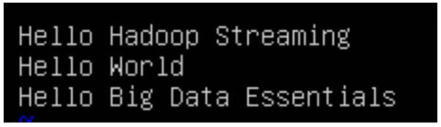
```
hadoopminhchau@minhchau-master:~$ echo "foo foo quux labs foo bar quux" | /home/hadoopminhchau/mappe
r.py
foo     1
foo     1
quux    1
labs    1
foo     1
bar     1
quux    1
hadoopminhchau@minhchau-master:~$
```

```
$ echo "foo foo quux labs foo bar quux" |
/home/hadoopminhchau/mapper.py | sort -k1,1 |
/home/hadoopminhchau/reducer.py
```



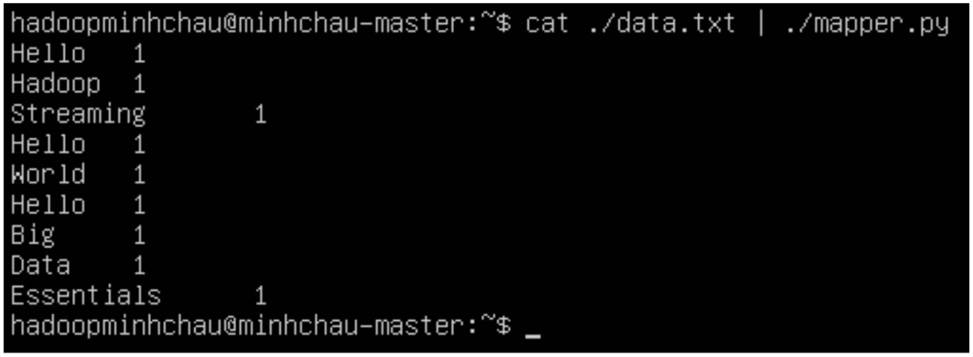
```
hadoopminhchau@minhchau-master:~$ echo "foo foo quux labs foo bar quux" | /home/hadoopminhchau/mappe
r.py
foo     1
foo     1
quux    1
labs    1
foo     1
bar     1
quux    1
hadoopminhchau@minhchau-master:~$ echo "foo foo quux labs foo bar quux" | /home/hadoopminhchau/mappe
r.py | sort -k1,1 | /home/hadoopminhchau/reducer.py
bar     1
foo     3
labs    1
quux    2
hadoopminhchau@minhchau-master:~$ _
```

- **Tạo file data.txt chứa dữ liệu**



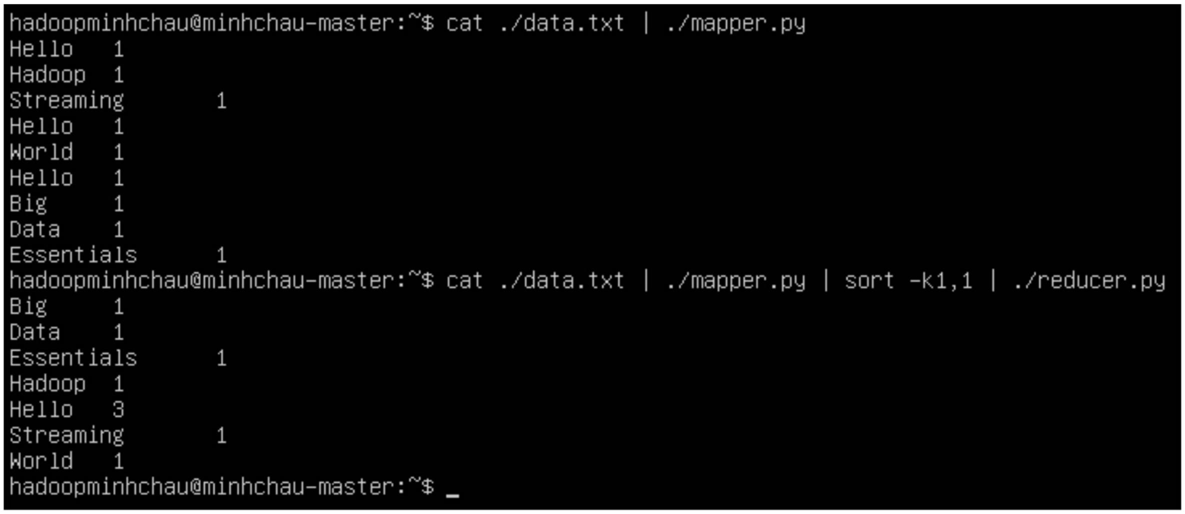
```
Hello Hadoop Streaming
Hello World
Hello Big Data Essentials
```

```
$ cat ./data.txt | ./mapper.py
```



```
hadoopminhchau@minhchau-master:~$ cat ./data.txt | ./mapper.py
Hello   1
Hadoop  1
Streaming       1
Hello   1
World   1
Hello   1
Big     1
Data    1
Essentials      1
hadoopminhchau@minhchau-master:~$ _
```

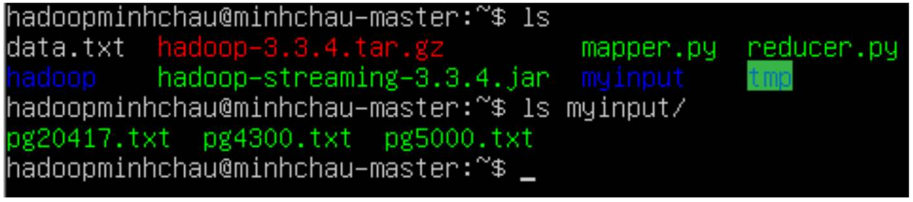
```
$ cat ./data.txt | ./mapper.py | sort -k1,1 | ./reducer.py
```



```
hadoopminhchau@minhchau-master:~$ cat ./data.txt | ./mapper.py
Hello   1
Hadoop  1
Streaming       1
Hello   1
World   1
Hello   1
Big     1
Data    1
Essentials      1
hadoopminhchau@minhchau-master:~$ cat ./data.txt | ./mapper.py | sort -k1,1 | ./reducer.py
Big     1
Data    1
Essentials      1
Hadoop  1
Hello   3
Streaming       1
World   1
hadoopminhchau@minhchau-master:~$ _
```

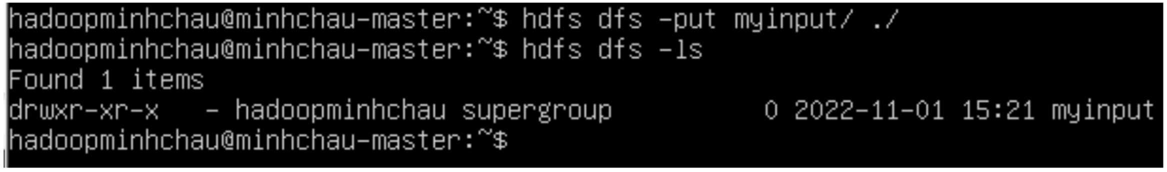
4. **Thực thi chương trình WordCount trên HDFS**

Tạo thư mục myinput chứa dữ liệu



```
hadoopminhchau@minhchau-master:~$ ls
data.txt  hadoop-3.3.4.tar.gz          mapper.py  reducer.py
hadoop    hadoop-streaming-3.3.4.jar  myinput    tmp
hadoopminhchau@minhchau-master:~$ ls myinput/
pg20417.txt  pg4300.txt  pg5000.txt
hadoopminhchau@minhchau-master:~$ _
```

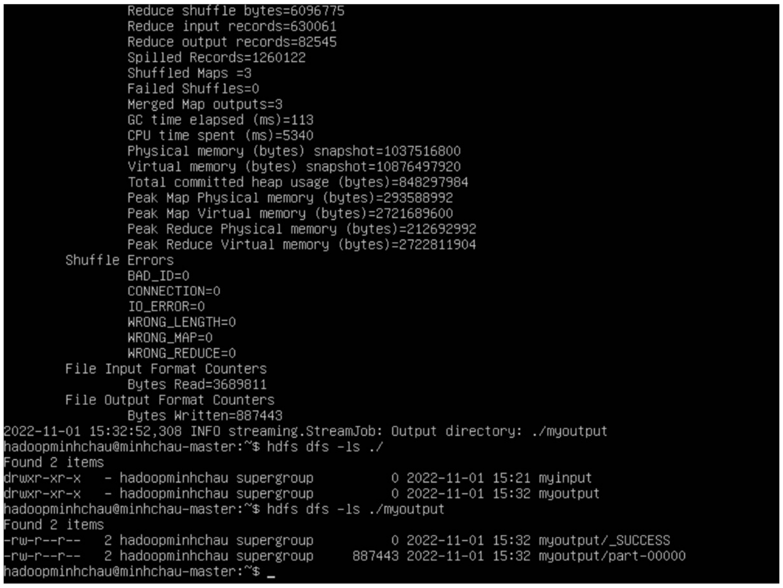
Copy thư mục myinput vào HDFS



```
hadoopminhchau@minhchau-master:~$ hdfs dfs -put myinput/ ./
hadoopminhchau@minhchau-master:~$ hdfs dfs -ls
Found 1 items
drwxr-xr-x   - hadoopminhchau supergroup          0 2022-11-01 15:21 myinput
hadoopminhchau@minhchau-master:~$
```

Chạy MapReduce job

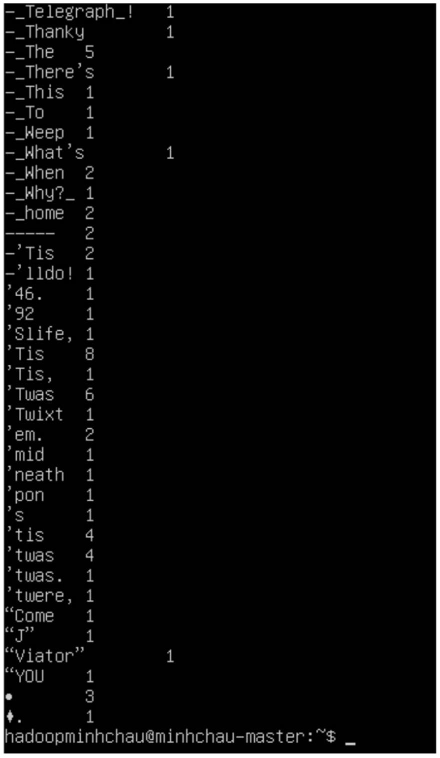
```
$ hadoop jar hadoop-streaming-3.3.4.jar -file mapper.py -
mapper mapper.py -file reducer.py -reducer reducer.py -
input ./myinput -output ./myoutput
```



```
                Reduce shuffle bytes=6096775
                Reduce input records=630061
                Reduce output records=82545
                Spilled Records=1260122
                Shuffled Maps =3
                Failed Shuffles=0
                Merged Map outputs=3
                GC time elapsed (ms)=113
                CPU time spent (ms)=5340
                Physical memory (bytes) snapshot=1037516800
                Virtual memory (bytes) snapshot=10876497920
                Total committed heap usage (bytes)=848297984
                Peak Map Physical memory (bytes)=293588992
                Peak Map Virtual memory (bytes)=2721689600
                Peak Reduce Physical memory (bytes)=212692992
                Peak Reduce Virtual memory (bytes)=2722811904
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=3689811
        File Output Format Counters
                Bytes Written=887443
2022-11-01 15:32:52,308 INFO streaming.StreamJob: Output directory: ./myoutput
hadoopminhchau@minhchau-master:~$ hdfs dfs -ls ./
Found 2 items
drwxr-xr-x   - hadoopminhchau supergroup          0 2022-11-01 15:21 myinput
drwxr-xr-x   - hadoopminhchau supergroup          0 2022-11-01 15:32 myoutput
hadoopminhchau@minhchau-master:~$ hdfs dfs -ls ./myoutput
Found 2 items
-rw-r--r--   2 hadoopminhchau supergroup          0 2022-11-01 15:32 myoutput/_SUCCESS
-rw-r--r--   2 hadoopminhchau supergroup     887443 2022-11-01 15:32 myoutput/part-00000
hadoopminhchau@minhchau-master:~$ _
```

Hiển thị kết quả

```
$ hdfs dfs -cat ./myoutput/part-00000
```



```
-_Telegraph_!   1
-_Thanky        1
-_The   5
-_There's       1
-_This  1
-_To    1
-_Weep  1
-_What's        1
-_When  2
-_Why?_ 1
-_home  2
-----   2
-'Tis   2
-'lldo! 1
'46.    1
'92     1
'Slife, 1
'Tis    8
'Tis,   1
'Twas   6
'Twixt  1
'em.    2
'mid    1
'neath  1
'pon    1
's      1
'tis    4
'twas   4
'twas.  1
'twere, 1
“Come   1
“J”     1
“Viator”        1
“YOU    1
•       3
♦.      1
hadoopminhchau@minhchau-master:~$ _
```

5. Sửa một số lỗi

Nếu báo lỗi/usr/bin/env: ‘python\r’: No such file or directory

```
$   sudo apt install dos2unix
```

Nếu báo lỗi /usr/bin/python^m bad interpreter

`$ vim mapper.py` then `:set ff=unix`

6. References

[1] https://www.tutorialspoint.com/hadoop/hadoop_streaming.htm

[2] https://www.tutsmake.com/how-to-install-python-3-10-on-ubuntu-22-04/

[3] https://www.michael-noll.com/tutorials/writing-an-hadoop-mapreduce-program-in-python/