1. Copy nhiều file từ máy thật sang máy ảo một cách nhanh nhất

   B1. Tạo thư mục chứa các file cần copy

   B2. Copy các file cần chuyển vào thư mục vừa tạo

   B3. Mở terminal tại thư mục vừa tạo

   B4. Gõ lệnh

   scp .* hduser@192.168.2.18:/home/hduser

```
PS C:\Users\thuan\Desktop\TÀI LIỆU ÔN THI CUỐI KÌ-20240415\MapReduce-Send-Student\MR-Send-Student\Map-reduce\Coppy> scp
.\* hduser@192.168.2.18:/home/hduser
hduser@192.168.2.18's password:
4300-0.txt                                    100% 1549KB  60.0MB/s   00:00
5000-8.txt                                    100% 1395KB  45.4MB/s   00:00
mapper.py                                     100%  566   362.6KB/s   00:00
pg20417.txt                                   100%  659KB  94.1MB/s   00:00
reducer.py                                    100% 1078     1.1MB/s   00:00
```

2. Tạo thư mục mycode:

   mkdir mycode

3. Di chuyển file code vào trong thư mục mycode:

   mv mapper.py mycode

   mv reducer.py mycode

4. Phân quyền 777 cho 2 file code:

   sudo chmod 777 mycode/reducer.py

   sudo chmod 777 mycode/mapper.py

```
hduser@osboxes ~ $ ls -l mycode
total 8
-rwxrwxrwx 1 hduser hadoop  566 Apr 22 06:28 mapper.py
-rwxrwxrwx 1 hduser hadoop 1078 Apr 22 06:28 reducer.py
hduser@osboxes ~ $
```

5. Chạy lệnh:

   start-all.sh

6. Tạo thư mục mydata trên HDFS:

   hdfs dfs -mkdir /mydata

7. Di chuyển các file từ local lên hdfs:

   hdfs dfs -copyFromLocal /path 1 /path 2 .... /path n /destination

   hdfs dfs -copyFromLocal /home/hduser/4300-0.txt /mydata

   hdfs dfs -copyFromLocal /home/hduser/5000-8.txt /mydata

   hdfs dfs -copyFromLocal /home/hduser/pg20417.txt /mydata

8. Lệnh kiểm tra xem đã copy được chưa:

   hdfs dfs -ls /mydata

```
hduser@osboxes ~ $ hdfs dfs -ls /mydata
Found 3 items
-rw-r--r--   1 hduser supergroup    1586336 2024-04-22 06:39 /mydata/4300-0.txt
-rw-r--r--   1 hduser supergroup    1428843 2024-04-22 06:39 /mydata/5000-8.txt
-rw-r--r--   1 hduser supergroup     674570 2024-04-22 06:39 /mydata/pg20417.txt
hduser@osboxes ~ $
```

9. Tạo thư mục myresult trên hdfs:

   hdfs dfs -mkdir /myresult

10. Kiểm tra thư mục được tạo thành công chưa:

    hdfs dfs -ls /

11. Kiểm tra bản python trên máy ảo: **whereis python** (Ta thấy có cả python v2 và v3)

```
hduser@osboxes ~ $ whereis python
python: /usr/bin/python /usr/bin/python3.5m /usr/bin/python2.7 /usr/bin/python3.5 /usr/lib/python2.7 /usr/lib/python2.6
/usr/lib/python3.5 /etc/python /etc/python2.7 /etc/python3.5 /usr/local/lib/python2.7 /usr/local/lib/python3.5 /usr/incl
ude/python3.5m /usr/include/python2.7 /usr/share/python /usr/share/man/man1/python.1.gz
hduser@osboxes ~ $
```

12. Kiểm tra xem python có chạy được mapper và reduce không:

    sudo echo "foo foo quiz lab foo bar quiz" | python /home/hduser/mycode/mapper.py

```
hduser@osboxes ~ $ sudo echo "foo foo quiz lab foo bar quiz" | python /home/hduser/mycode/mapper.py
foo     1
foo     1
quiz    1
lab     1
foo     1
bar     1
quiz    1
hduser@osboxes ~ $
```

13. Kiểm tra có Hadoop stream hay chưa:

    cd $HADOOP_HOME

    ls -l share/hadoop/tools/lib/

```
-rw-r--r-- 1 hduser hadoop    37985 Oct  3  2016 hadoop-datajoin-2.6.5.jar
-rw-r--r-- 1 hduser hadoop   117477 Oct  3  2016 hadoop-distcp-2.6.5.jar
-rw-r--r-- 1 hduser hadoop    85426 Oct  3  2016 hadoop-extras-2.6.5.jar
-rw-r--r-- 1 hduser hadoop   238266 Oct  3  2016 hadoop-gridmix-2.6.5.jar
-rw-r--r-- 1 hduser hadoop   136950 Oct  3  2016 hadoop-openstack-2.6.5.jar
-rw-r--r-- 1 hduser hadoop   301108 Oct  3  2016 hadoop-rumen-2.6.5.jar
-rw-r--r-- 1 hduser hadoop   137426 Oct  3  2016 hadoop-sls-2.6.5.jar
-rw-r--r-- 1 hduser hadoop   128286 Oct  3  2016 hadoop-streaming-2.6.5.jar
-rw-r--r-- 1 hduser hadoop    45024 Oct  3  2016 hamcrest-core-1.3.jar
```

14. Sử dụng Hadoop stream để chạy map-reduce đếm số:

hadoop jar share/hadoop/tools/lib/hadoop-streaming-2.6.5.jar -files "/home/hduser/mycode/mapper.py,/home/hduser/mycode/reducer.py" -mapper "python mapper.py" -reducer "python reducer.py" -input /mydata/* -output /myresult/out-res01

```
        File Output Format Counters
                Bytes Written=887415
24/04/22 06:54:12 INFO streaming.StreamJob: Output directory: /myresult/out-res01
hduser@osboxes /usr/local/hadoop $ hadoop jar share/hadoop/tools/lib/hadoop-streaming-2.6.5.jar -files "/home/hduser/myc
ode/mapper.py,/home/hduser/mycode/reducer.py" -mapper "python mapper.py" -reducer "python reducer.py" -input /mydata/* -
output /myresult/out-res01
```

15. Kiểm tra file đã tạo thành công chưa:

    hdfs dfs -ls /myresult/out-res01

```
hduser@osboxes /usr/local/hadoop $ hdfs dfs -ls /myresult/out-res01
Found 2 items
-rw-r--r--   1 hduser supergroup          0 2024-04-22 06:54 /myresult/out-res01/_SUCCESS
-rw-r--r--   1 hduser supergroup     887415 2024-04-22 06:54 /myresult/out-res01/part-00000
hduser@osboxes /usr/local/hadoop $
```

16. Đọc file vừa được tạo ra:

    hdfs dfs -cat /myresult/out-res01/part-00000

```
“Viator”        1
“YOU    1
•       1
✠       1
       3
:       1
crit_   1
hduser@osboxes /usr/local/hadoop $ hdfs dfs -cat /myresult/out-res01/part-00000
```

CHẠY TEST BÊN WINDOW

1. Viết 2 file mapper và reduce bên windowns
2. Mở Project bằng vs code có chưa cả CSV để gợi ý code chính xác hơn
3. Chạy test code bằng lệnh
   Chạy mapper:
   python mapper.py < Path_CSV
   Chạy map-reduce:
   python mapper.py < Path_CSV | python reducer.py

```
C:\Users\thuan\Desktop\Map-reduce_example\code01\01>python mapper.py < C:\Users\thuan\Desktop\Map-reduce_example\code01\emp_data.csv | python reducer.py
min     800

C:\Users\thuan\Desktop\Map-reduce_example\code01\01>
```

4. Sau khi test hết chức năng thì tiến hành copy file data và code như ở phần hướng dẫn trên