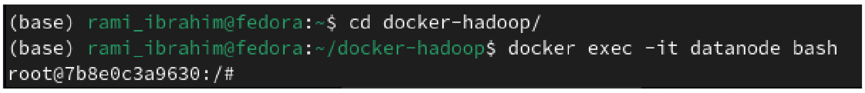
> Rami Ibrahim ING 2 TD 2
> Compte rendu TP2 BIG DATA

In this TP we'll go over Hadoop Streaming, so we need some dependencies that we need to install in order to kick start this , let's do it together !

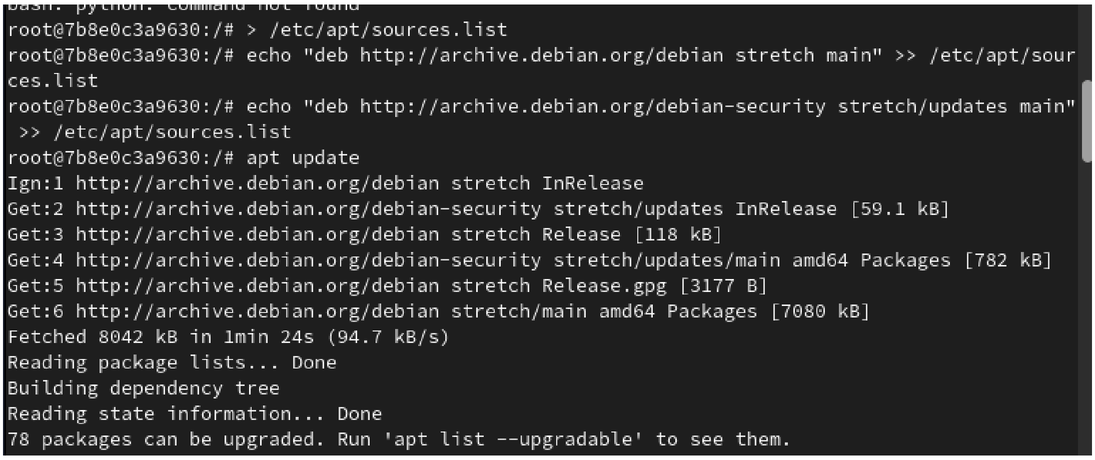
Let's start by installing `python3` on these nodes

Let's open our terminal , move inside of the directory and execute bash on our datanode



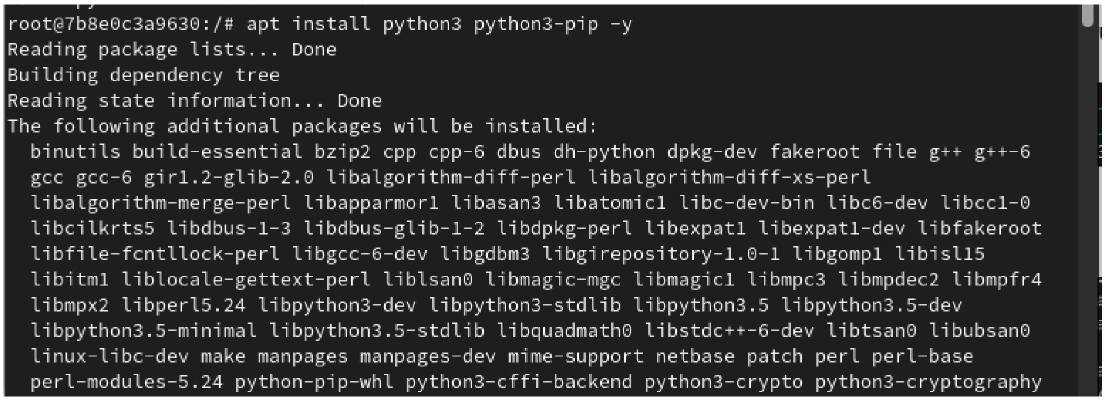
```
(base) rami_ibrahim@fedora:~$ cd docker-hadoop/
(base) rami_ibrahim@fedora:~/docker-hadoop$ docker exec -it datanode bash
root@7b8e0c3a9630:/#
```

To install python3 we need some required reposotries, let's add them



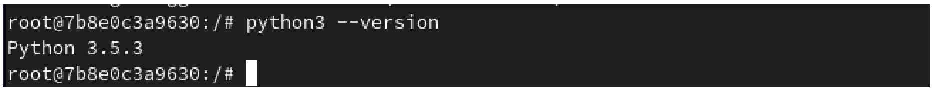
```
bash: python: command not found
root@7b8e0c3a9630:/# > /etc/apt/sources.list
root@7b8e0c3a9630:/# echo "deb http://archive.debian.org/debian stretch main" >> /etc/apt/sour
ces.list
root@7b8e0c3a9630:/# echo "deb http://archive.debian.org/debian-security stretch/updates main"
 >> /etc/apt/sources.list
root@7b8e0c3a9630:/# apt update
Ign:1 http://archive.debian.org/debian stretch InRelease
Get:2 http://archive.debian.org/debian-security stretch/updates InRelease [59.1 kB]
Get:3 http://archive.debian.org/debian stretch Release [118 kB]
Get:4 http://archive.debian.org/debian-security stretch/updates/main amd64 Packages [782 kB]
Get:5 http://archive.debian.org/debian stretch Release.gpg [3177 B]
Get:6 http://archive.debian.org/debian stretch/main amd64 Packages [7080 kB]
Fetched 8042 kB in 1min 24s (94.7 kB/s)
Reading package lists... Done
Building dependency tree
Reading state information... Done
78 packages can be upgraded. Run 'apt list --upgradable' to see them.
```

then update the packages and run the commande to install python3



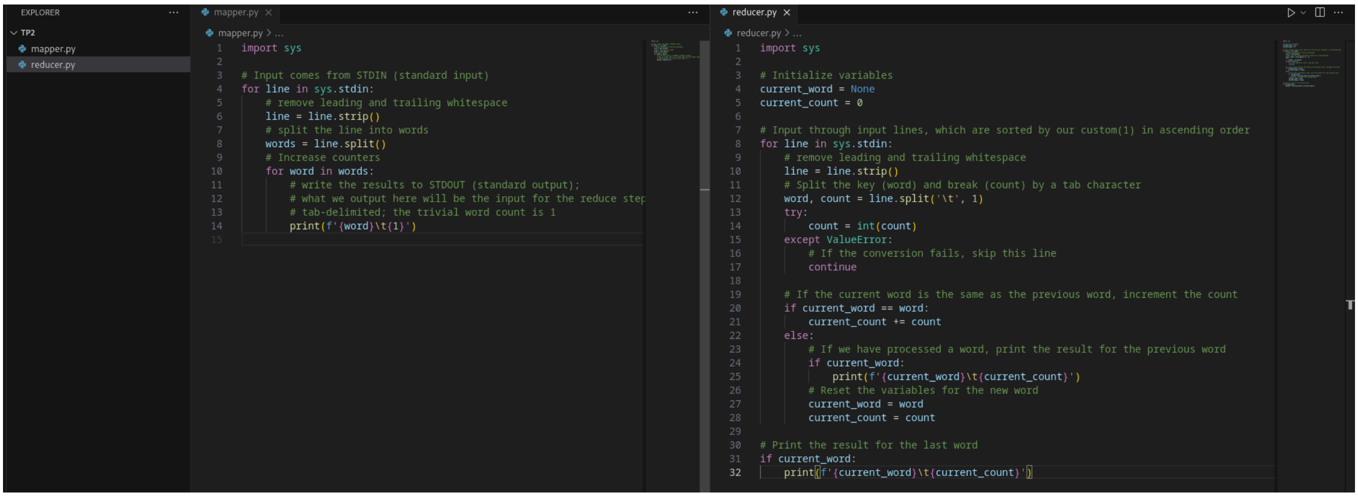
```
root@7b8e0c3a9630:/# apt install python3 python3-pip -y
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following additional packages will be installed:
  binutils build-essential bzip2 cpp cpp-6 dbus dh-python dpkg-dev fakeroot file g++ g++-6
  gcc gcc-6 gir1.2-glib-2.0 libalgorithm-diff-perl libalgorithm-diff-xs-perl
  libalgorithm-merge-perl libapparmor1 libasan3 libatomic1 libc-dev-bin libc6-dev libcc1-0
  libcilkrts5 libdbus-1-3 libdbus-glib-1-2 libdpkg-perl libexpat1 libexpat1-dev libfakeroot
  libfile-fcntllock-perl libgcc-6-dev libgdbm3 libgirepository-1.0-1 libgomp1 libisl15
  libitm1 liblocale-gettext-perl liblsan0 libmagic-mgc libmagic1 libmpc3 libmpdec2 libmpfr4
  libmpx2 libperl5.24 libpython3-dev libpython3-stdlib libpython3.5 libpython3.5-dev
  libpython3.5-minimal libpython3.5-stdlib libquadmath0 libstdc++-6-dev libtsan0 libubsan0
  linux-libc-dev make manpages manpages-dev mime-support netbase patch perl perl-base
  perl-modules-5.24 python-pip-whl python3-cffi-backend python3-crypto python3-cryptography
```

Everything seems fine, let's check the installation



```
root@7b8e0c3a9630:/# python3 --version
Python 3.5.3
root@7b8e0c3a9630:/#
```

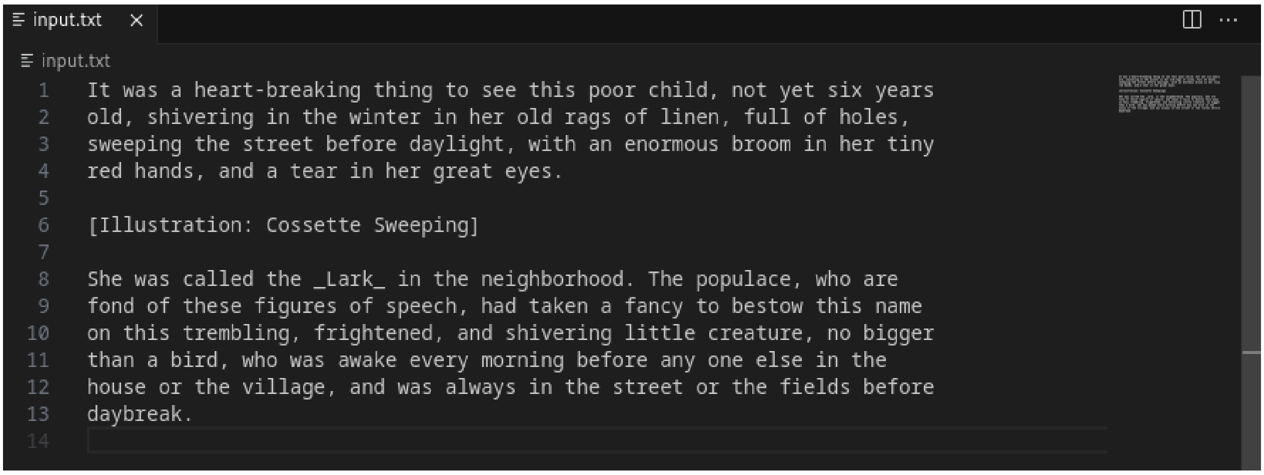
Once we finish the installation let's do some actual coding, we'll be needing 2 python files: `mapper.py` and `reducer.py`



```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
        # tab-delimited; the trivial word count is 1
        print(f'{word}\t{1}')
```

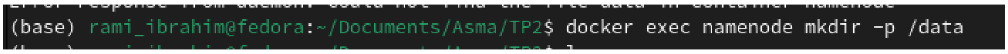
```python
import sys

# Initialize variables
current_word = None
current_count = 0

# Input through input lines, which are sorted by our custom(1) in ascending order
for line in sys.stdin:
    # remove leading and trailing whitespace
    line = line.strip()
    # Split the key (word) and break (count) by a tab character
    word, count = line.split('\t', 1)
    try:
        count = int(count)
    except ValueError:
        # If the conversion fails, skip this line
        continue

    # If the current word is the same as the previous word, increment the count
    if current_word == word:
        current_count += count
    else:
        # If we have processed a word, print the result for the previous word
        if current_word:
            print(f'{current_word}\t{current_count}')
        # Reset the variables for the new word
        current_word = word
        current_count = count

# Print the result for the last word
if current_word:
    print(f'{current_word}\t{current_count}')
```

Create a text file `input.txt



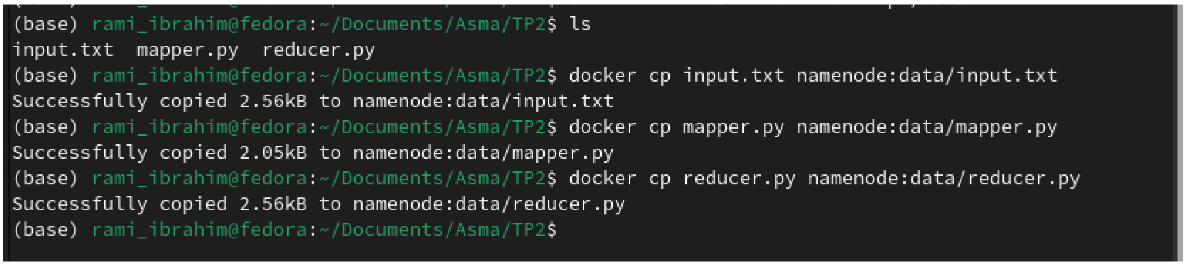
```
It was a heart-breaking thing to see this poor child, not yet six years
old, shivering in the winter in her old rags of linen, full of holes,
sweeping the street before daylight, with an enormous broom in her tiny
red hands, and a tear in her great eyes.

[Illustration: Cossette Sweeping]

She was called the _Lark_ in the neighborhood. The populace, who are
fond of these figures of speech, had taken a fancy to bestow this name
on this trembling, frightened, and shivering little creature, no bigger
than a bird, who was awake every morning before any one else in the
house or the village, and was always in the street or the fields before
daybreak.
```

Go inside the container and create a folder named `data`



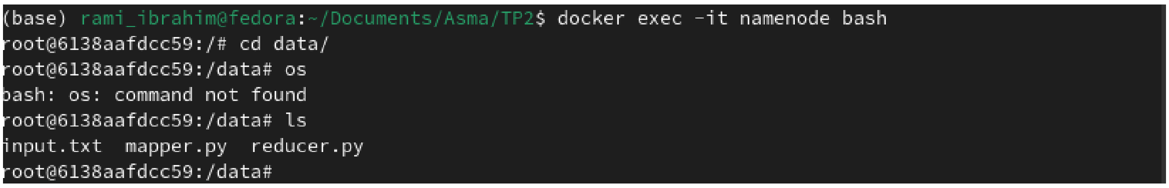
```
(base) rami_ibrahim@fedora:~/Documents/Asma/TP2$ docker exec namenode mkdir -p /data
```

Let's move these files into the container (from native)



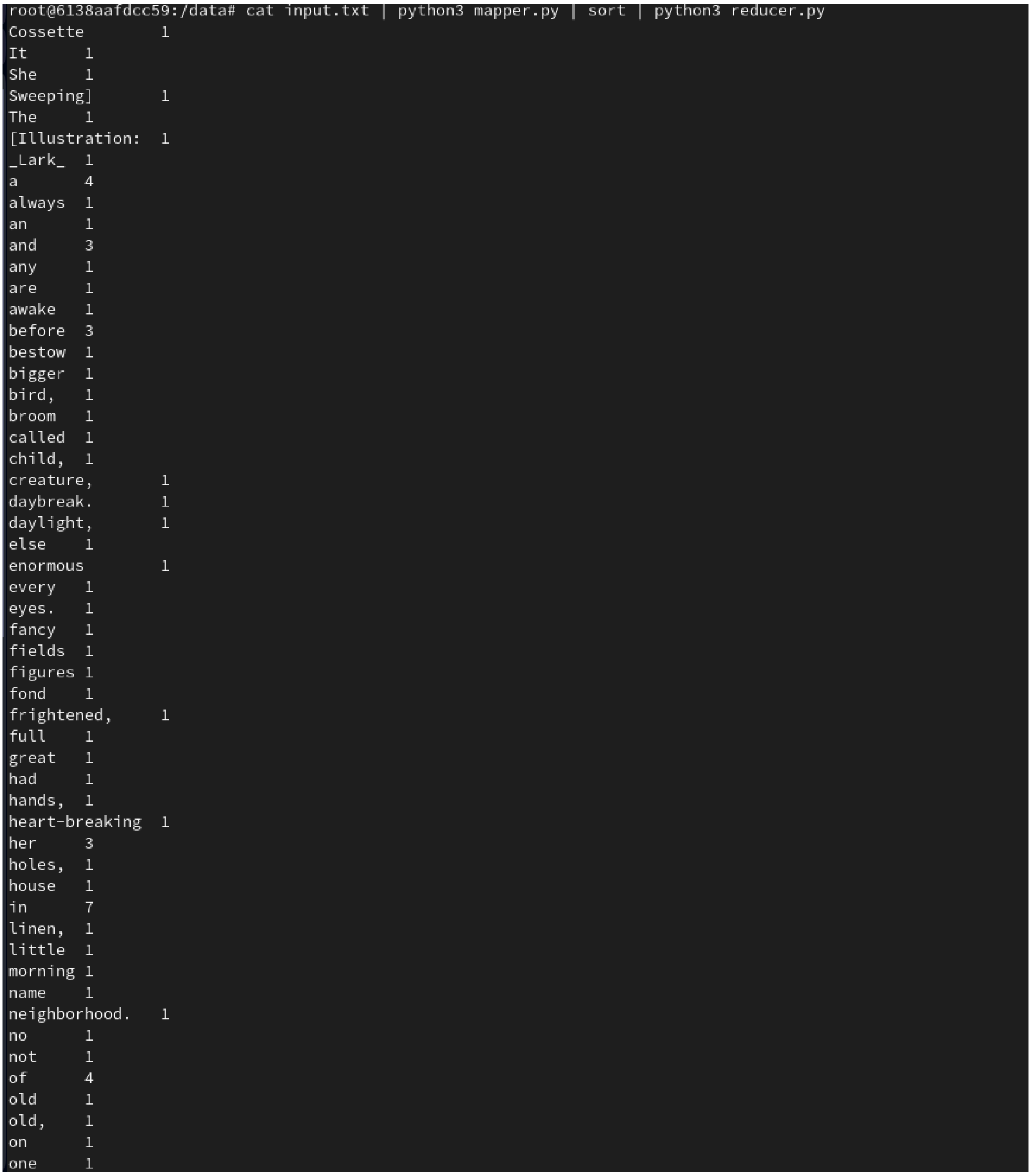
```
(base) rami_ibrahim@fedora:~/Documents/Asma/TP2$ ls
input.txt  mapper.py  reducer.py
(base) rami_ibrahim@fedora:~/Documents/Asma/TP2$ docker cp input.txt namenode:data/input.txt
Successfully copied 2.56kB to namenode:data/input.txt
(base) rami_ibrahim@fedora:~/Documents/Asma/TP2$ docker cp mapper.py namenode:data/mapper.py
Successfully copied 2.05kB to namenode:data/mapper.py
(base) rami_ibrahim@fedora:~/Documents/Asma/TP2$ docker cp reducer.py namenode:data/reducer.py
Successfully copied 2.56kB to namenode:data/reducer.py
(base) rami_ibrahim@fedora:~/Documents/Asma/TP2$
```

Open the container bash and verify



```
(base) rami_ibrahim@fedora:~/Documents/Asma/TP2$ docker exec -it namenode bash
root@6138aafdcc59:/# cd data/
root@6138aafdcc59:/data# os
bash: os: command not found
root@6138aafdcc59:/data# ls
input.txt  mapper.py  reducer.py
root@6138aafdcc59:/data#
```

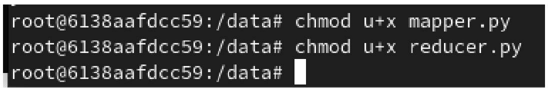
and our files are copied, let's execute them !



```
root@6138aafdcc59:/data# cat input.txt | python3 mapper.py | sort | python3 reducer.py
Cossette        1
It      1
She     1
Sweeping]       1
The     1
[Illustration:  1
_Lark_  1
a       4
always  1
an      1
and     3
any     1
are     1
awake   1
before  3
bestow  1
bigger  1
bird,   1
broom   1
called  1
child,  1
creature,       1
daybreak.       1
daylight,       1
else    1
enormous        1
every   1
eyes.   1
fancy   1
fields  1
figures 1
fond    1
frightened,     1
full    1
great   1
had     1
hands,  1
heart-breaking  1
her     3
holes,  1
house   1
in      7
linen,  1
little  1
morning 1
name    1
neighborhood.   1
no      1
not     1
of      4
old     1
old,    1
on      1
one     1
```

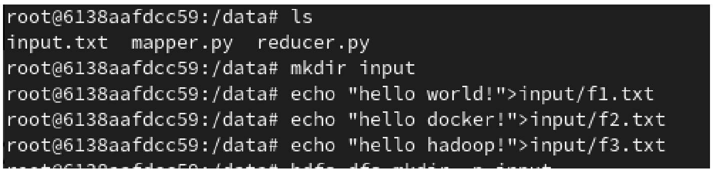
Now let's execute this inside of the HDFS

First, let's change the mode of the python files

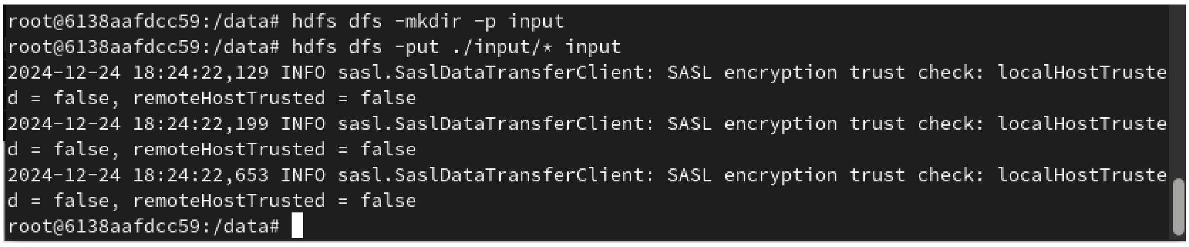


```
root@6138aafdcc59:/data# chmod u+x mapper.py
root@6138aafdcc59:/data# chmod u+x reducer.py
root@6138aafdcc59:/data# █
```

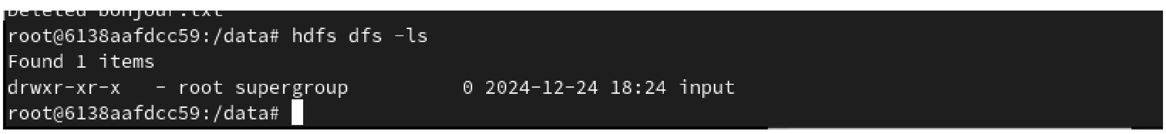
Now create the files and transfer them to HDFS



```
root@6138aafdcc59:/data# ls
input.txt  mapper.py  reducer.py
root@6138aafdcc59:/data# mkdir input
root@6138aafdcc59:/data# echo "hello world!">input/f1.txt
root@6138aafdcc59:/data# echo "hello docker!">input/f2.txt
root@6138aafdcc59:/data# echo "hello hadoop!">input/f3.txt
root@6138aafdcc59:/data# hdfs dfs -mkdir -p input
```

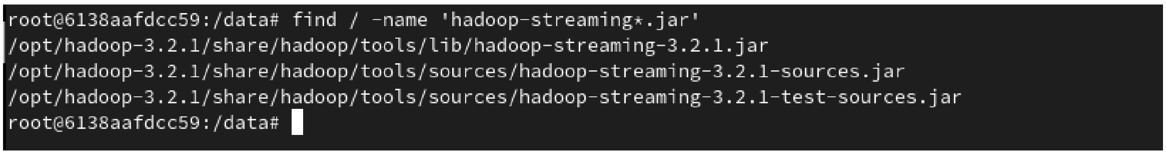


```
root@6138aafdcc59:/data# hdfs dfs -mkdir -p input
root@6138aafdcc59:/data# hdfs dfs -put ./input/* input
2024-12-24 18:24:22,129 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTruste
d = false, remoteHostTrusted = false
2024-12-24 18:24:22,199 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTruste
d = false, remoteHostTrusted = false
2024-12-24 18:24:22,653 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTruste
d = false, remoteHostTrusted = false
root@6138aafdcc59:/data#
```

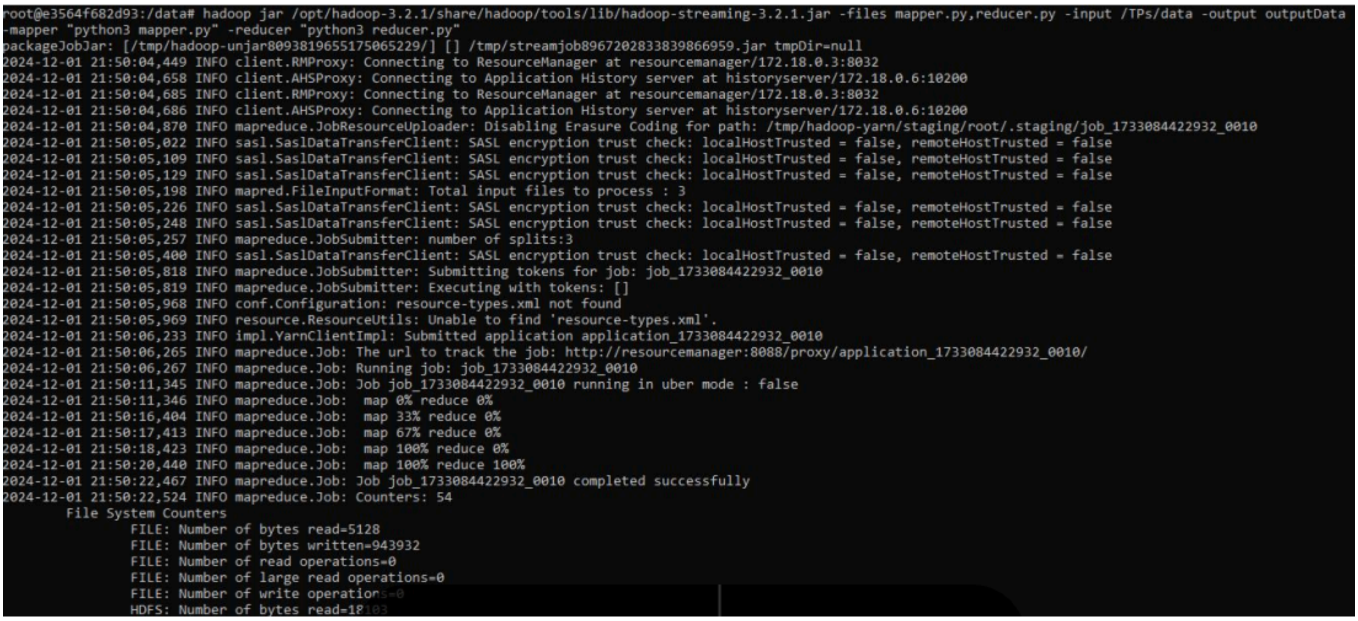


```
Deleted bonjour.txt
root@6138aafdcc59:/data# hdfs dfs -ls
Found 1 items
drwxr-xr-x   - root supergroup          0 2024-12-24 18:24 input
root@6138aafdcc59:/data#
```

Now execute the program MapReduce

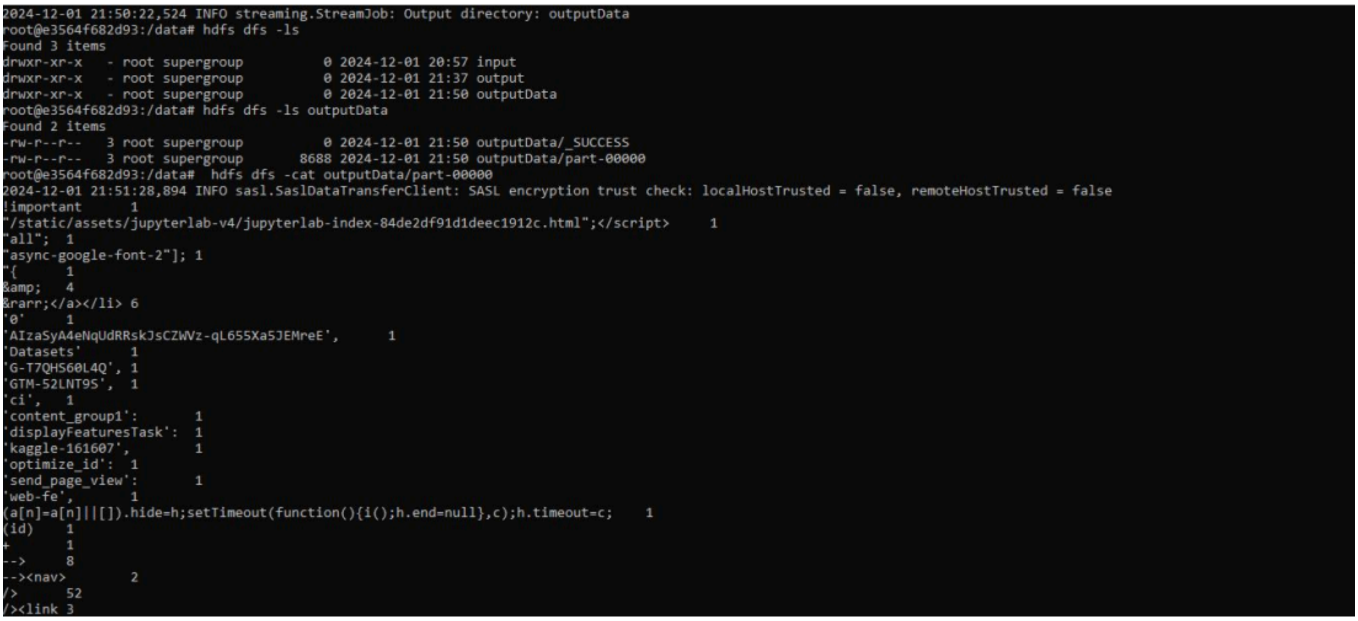


```
root@6138aafdcc59:/data# find / -name 'hadoop-streaming*.jar'
/opt/hadoop-3.2.1/share/hadoop/tools/lib/hadoop-streaming-3.2.1.jar
/opt/hadoop-3.2.1/share/hadoop/tools/sources/hadoop-streaming-3.2.1-sources.jar
/opt/hadoop-3.2.1/share/hadoop/tools/sources/hadoop-streaming-3.2.1-test-sources.jar
root@6138aafdcc59:/data#
```



```
root@e3564f682d93:/data# hadoop jar /opt/hadoop-3.2.1/share/hadoop/tools/lib/hadoop-streaming-3.2.1.jar -files mapper.py,reducer.py -input /TPs/data -output outputData
-mapper "python3 mapper.py" -reducer "python3 reducer.py"
packageJobJar: [/tmp/hadoop-unjar8093819655175065229/] [] /tmp/streamjob8967202833839866959.jar tmpDir=null
2024-12-01 21:50:04,449 INFO client.RMProxy: Connecting to ResourceManager at resourcemanager/172.18.0.3:8032
2024-12-01 21:50:04,658 INFO client.AHSProxy: Connecting to Application History server at historyserver/172.18.0.6:10200
2024-12-01 21:50:04,685 INFO client.RMProxy: Connecting to ResourceManager at resourcemanager/172.18.0.3:8032
2024-12-01 21:50:04,686 INFO client.AHSProxy: Connecting to Application History server at historyserver/172.18.0.6:10200
2024-12-01 21:50:04,870 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1733084422932_0010
2024-12-01 21:50:05,022 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-12-01 21:50:05,109 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-12-01 21:50:05,129 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-12-01 21:50:05,198 INFO mapred.FileInputFormat: Total input files to process : 3
2024-12-01 21:50:05,226 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-12-01 21:50:05,248 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-12-01 21:50:05,257 INFO mapreduce.JobSubmitter: number of splits:3
2024-12-01 21:50:05,400 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-12-01 21:50:05,818 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1733084422932_0010
2024-12-01 21:50:05,819 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-12-01 21:50:05,968 INFO conf.Configuration: resource-types.xml not found
2024-12-01 21:50:05,969 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-12-01 21:50:06,233 INFO impl.YarnClientImpl: Submitted application application_1733084422932_0010
2024-12-01 21:50:06,265 INFO mapreduce.Job: The url to track the job: http://resourcemanager:8088/proxy/application_1733084422932_0010/
2024-12-01 21:50:06,267 INFO mapreduce.Job: Running job: job_1733084422932_0010
2024-12-01 21:50:11,345 INFO mapreduce.Job: Job job_1733084422932_0010 running in uber mode : false
2024-12-01 21:50:11,346 INFO mapreduce.Job:  map 0% reduce 0%
2024-12-01 21:50:16,404 INFO mapreduce.Job:  map 33% reduce 0%
2024-12-01 21:50:17,413 INFO mapreduce.Job:  map 67% reduce 0%
2024-12-01 21:50:18,423 INFO mapreduce.Job:  map 100% reduce 0%
2024-12-01 21:50:20,440 INFO mapreduce.Job:  map 100% reduce 100%
2024-12-01 21:50:22,467 INFO mapreduce.Job: Job job_1733084422932_0010 completed successfully
2024-12-01 21:50:22,524 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=5128
                FILE: Number of bytes written=943932
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=18103
```

And view the result



```
2024-12-01 21:50:22,524 INFO streaming.StreamJob: Output directory: outputData
root@e3564f682d93:/data# hdfs dfs -ls
Found 3 items
drwxr-xr-x   - root supergroup          0 2024-12-01 20:57 input
drwxr-xr-x   - root supergroup          0 2024-12-01 21:37 output
drwxr-xr-x   - root supergroup          0 2024-12-01 21:50 outputData
root@e3564f682d93:/data# hdfs dfs -ls outputData
Found 2 items
-rw-r--r--   3 root supergroup          0 2024-12-01 21:50 outputData/_SUCCESS
-rw-r--r--   3 root supergroup       8688 2024-12-01 21:50 outputData/part-00000
root@e3564f682d93:/data#  hdfs dfs -cat outputData/part-00000
2024-12-01 21:51:28,894 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
!important      1
"/static/assets/jupyterlab-v4/jupyterlab-index-84de2df91d1deec1912c.html";</script>     1
"all";  1
"async-google-font-2"]; 1
"{      1
&amp;   4
&rarr;</a></li> 6
'0'     1
'AIzaSyA4eNqUdRRskJsCZWVz-qL655Xa5JEMreE',      1
'Datasets'      1
'G-T7QHS60L4Q', 1
'GTM-52LNT9S',  1
'ci',   1
'content_group1':       1
'displayFeaturesTask':  1
'kaggle-161607',        1
'optimize_id':  1
'send_page_view':       1
'web-fe',       1
(a[n]=a[n]||[]).hide=h;setTimeout(function(){i();h.end=null},c);h.timeout=c;    1
(id)    1
+       1
-->     8
--><nav>        2
/>      52
/><link 3
```