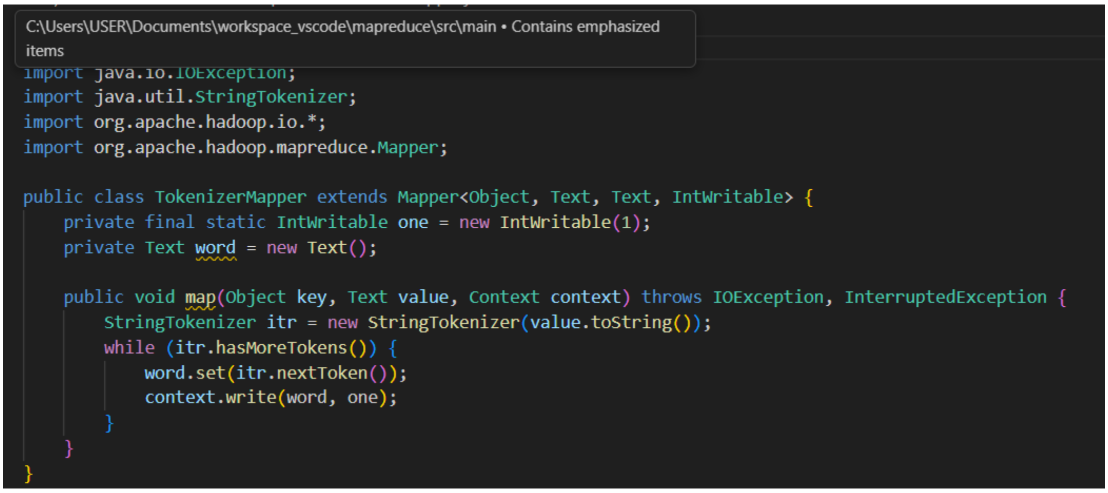
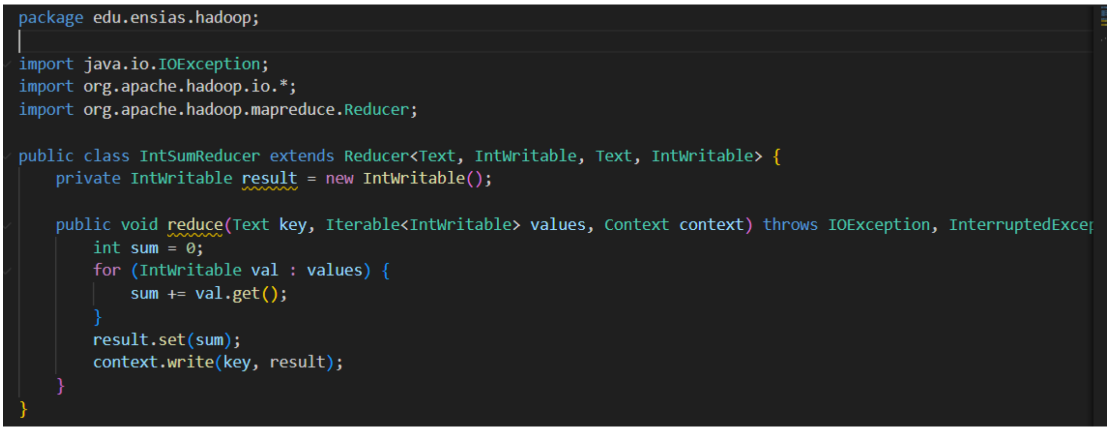
TP : programmation avec l’API MapReduce

1)classe Mapper



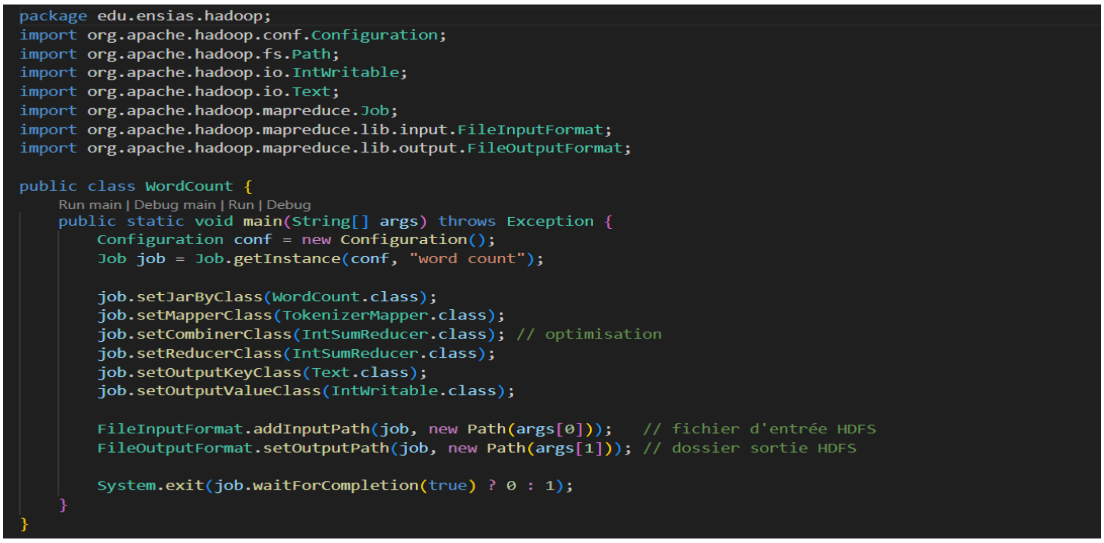
```
C:\Users\USER\Documents\workspace_vscode\mapreduce\src\main • Contains emphasized items
import java.io.IOException;
import java.util.StringTokenizer;
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapreduce.Mapper;

public class TokenizerMapper extends Mapper<Object, Text, Text, IntWritable> {
    private final static IntWritable one = new IntWritable(1);
    private Text word = new Text();

    public void map(Object key, Text value, Context context) throws IOException, InterruptedException {
        StringTokenizer itr = new StringTokenizer(value.toString());
        while (itr.hasMoreTokens()) {
            word.set(itr.nextToken());
            context.write(word, one);
        }
    }
}
```

2)classe reducer



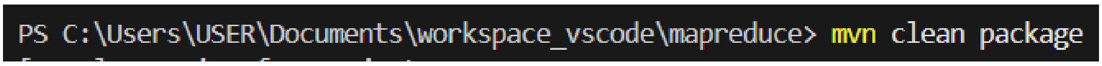
```
package edu.ensias.hadoop;

import java.io.IOException;
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapreduce.Reducer;

public class IntSumReducer extends Reducer<Text, IntWritable, Text, IntWritable> {
    private IntWritable result = new IntWritable();

    public void reduce(Text key, Iterable<IntWritable> values, Context context) throws IOException, InterruptedExcep
        int sum = 0;
        for (IntWritable val : values) {
            sum += val.get();
        }
        result.set(sum);
        context.write(key, result);
    }
}
```

3)Main classe



```
package edu.ensias.hadoop;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

public class WordCount {
    Run main | Debug main | Run | Debug
    public static void main(String[] args) throws Exception {
        Configuration conf = new Configuration();
        Job job = Job.getInstance(conf, "word count");

        job.setJarByClass(WordCount.class);
        job.setMapperClass(TokenizerMapper.class);
        job.setCombinerClass(IntSumReducer.class); // optimisation
        job.setReducerClass(IntSumReducer.class);
        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(IntWritable.class);

        FileInputFormat.addInputPath(job, new Path(args[0]));   // fichier d'entrée HDFS
        FileOutputFormat.setOutputPath(job, new Path(args[1])); // dossier sortie HDFS

        System.exit(job.waitForCompletion(true) ? 0 : 1);
    }
}
```

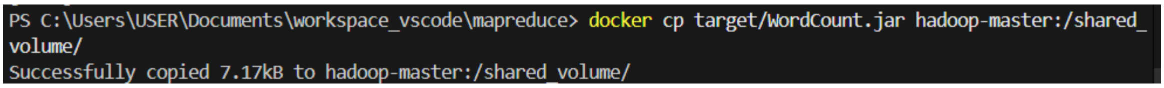
Créer un fichier jar que vous allez nommé WordCount.jar



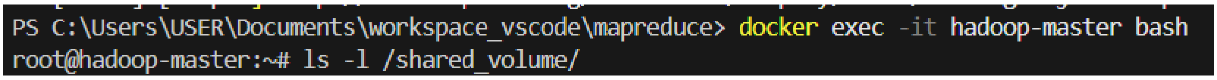
```
PS C:\Users\USER\Documents\workspace_vscode\mapreduce> mvn clean package
```

Building jar: C:\Users\USER\Documents\workspace_vscode\mapreduce\target\WordCount.jar

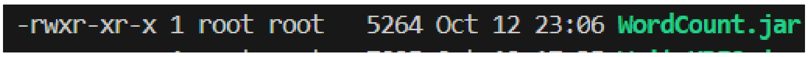
Copier le jar créé vers le dossier de partage /hadoop_project



```
PS C:\Users\USER\Documents\workspace_vscode\mapreduce> docker cp target/WordCount.jar hadoop-master:/shared_
volume/
Successfully copied 7.17kB to hadoop-master:/shared_volume/
```

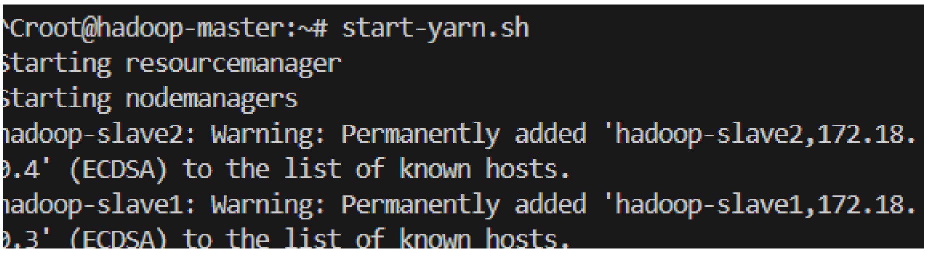


```
PS C:\Users\USER\Documents\workspace_vscode\mapreduce> docker exec -it hadoop-master bash
root@hadoop-master:~# ls -l /shared_volume/
```

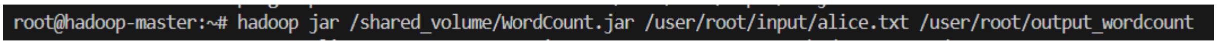


```
-rwxr-xr-x 1 root root   5264 Oct 12 23:06 WordCount.jar
```

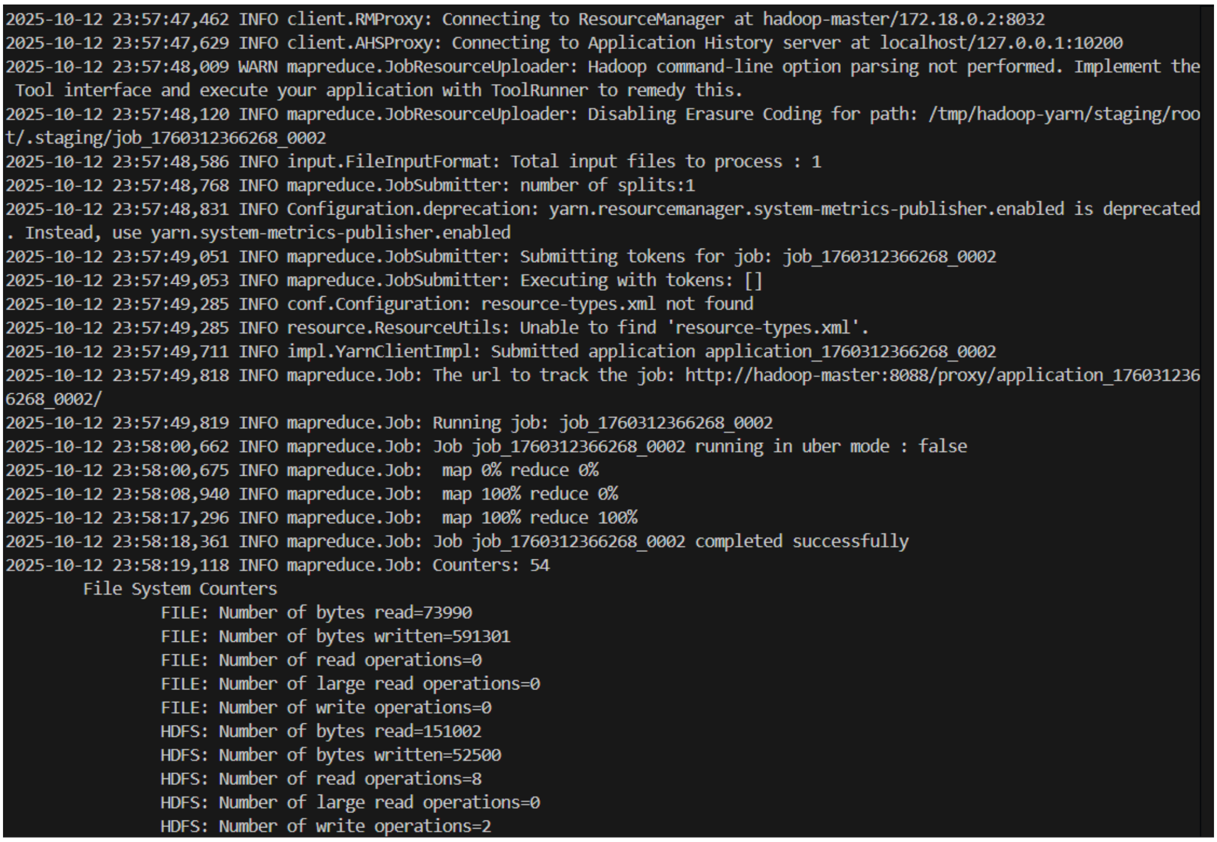
Sur l’invité de commande shell de votre container lancer la commande



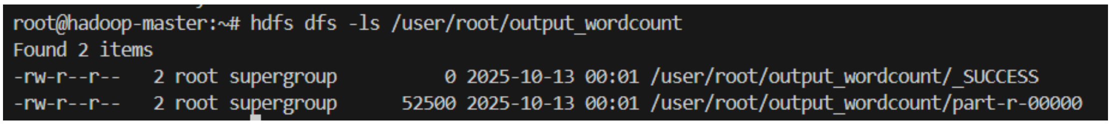
```
^Croot@hadoop-master:~# start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hadoop-slave2: Warning: Permanently added 'hadoop-slave2,172.18.
0.4' (ECDSA) to the list of known hosts.
hadoop-slave1: Warning: Permanently added 'hadoop-slave1,172.18.
0.3' (ECDSA) to the list of known hosts.
```



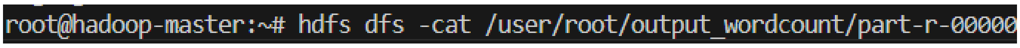
```
root@hadoop-master:~# hadoop jar /shared_volume/WordCount.jar /user/root/input/alice.txt /user/root/output_wordcount
```



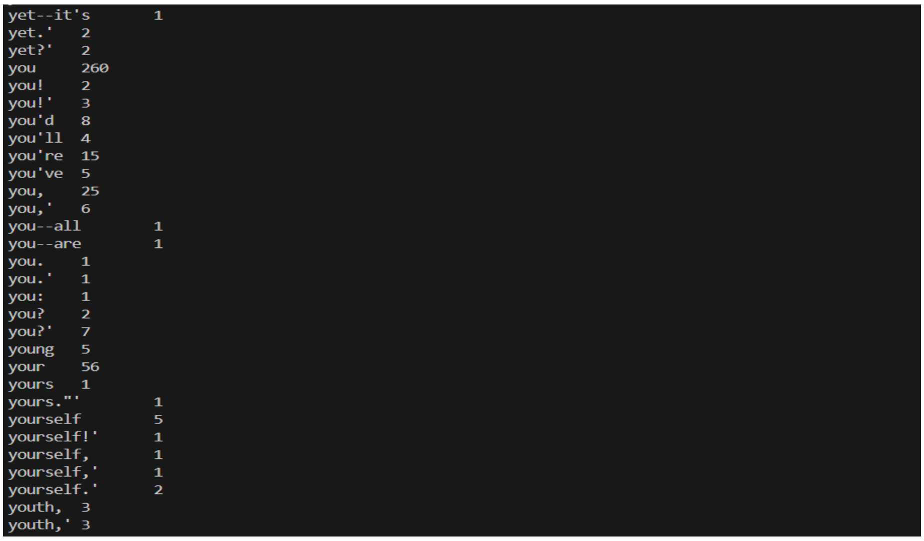
```
2025-10-12 23:57:47,462 INFO client.RMProxy: Connecting to ResourceManager at hadoop-master/172.18.0.2:8032
2025-10-12 23:57:47,629 INFO client.AHSProxy: Connecting to Application History server at localhost/127.0.0.1:10200
2025-10-12 23:57:48,009 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the
 Tool interface and execute your application with ToolRunner to remedy this.
2025-10-12 23:57:48,120 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/roo
t/.staging/job_1760312366268_0002
2025-10-12 23:57:48,586 INFO input.FileInputFormat: Total input files to process : 1
2025-10-12 23:57:48,768 INFO mapreduce.JobSubmitter: number of splits:1
2025-10-12 23:57:48,831 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated
. Instead, use yarn.system-metrics-publisher.enabled
2025-10-12 23:57:49,051 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1760312366268_0002
2025-10-12 23:57:49,053 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-10-12 23:57:49,285 INFO conf.Configuration: resource-types.xml not found
2025-10-12 23:57:49,285 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-10-12 23:57:49,711 INFO impl.YarnClientImpl: Submitted application application_1760312366268_0002
2025-10-12 23:57:49,818 INFO mapreduce.Job: The url to track the job: http://hadoop-master:8088/proxy/application_176031236
6268_0002/
2025-10-12 23:57:49,819 INFO mapreduce.Job: Running job: job_1760312366268_0002
2025-10-12 23:58:00,662 INFO mapreduce.Job: Job job_1760312366268_0002 running in uber mode : false
2025-10-12 23:58:00,675 INFO mapreduce.Job:  map 0% reduce 0%
2025-10-12 23:58:08,940 INFO mapreduce.Job:  map 100% reduce 0%
2025-10-12 23:58:17,296 INFO mapreduce.Job:  map 100% reduce 100%
2025-10-12 23:58:18,361 INFO mapreduce.Job: Job job_1760312366268_0002 completed successfully
2025-10-12 23:58:19,118 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=73990
                FILE: Number of bytes written=591301
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=151002
                HDFS: Number of bytes written=52500
                HDFS: Number of read operations=8
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
```



```
root@hadoop-master:~# hdfs dfs -ls /user/root/output_wordcount
Found 2 items
-rw-r--r--   2 root supergroup          0 2025-10-13 00:01 /user/root/output_wordcount/_SUCCESS
-rw-r--r--   2 root supergroup      52500 2025-10-13 00:01 /user/root/output_wordcount/part-r-00000
```



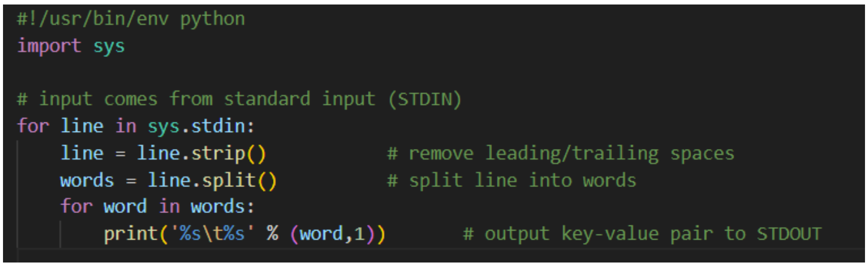
```
root@hadoop-master:~# hdfs dfs -cat /user/root/output_wordcount/part-r-00000
```



```
yet--it's       1
yet.'   2
yet?'   2
you     260
you!    2
you!'   3
you'd   8
you'll  4
you're  15
you've  5
you,    25
you,'   6
you--all        1
you--are        1
you.    1
you.'   1
you:    1
you?    2
you?'   7
young   5
your    56
yours   1
yours."'        1
yourself        5
yourself!'      1
yourself,       1
yourself,'      1
yourself.'      2
youth,  3
youth,' 3
```

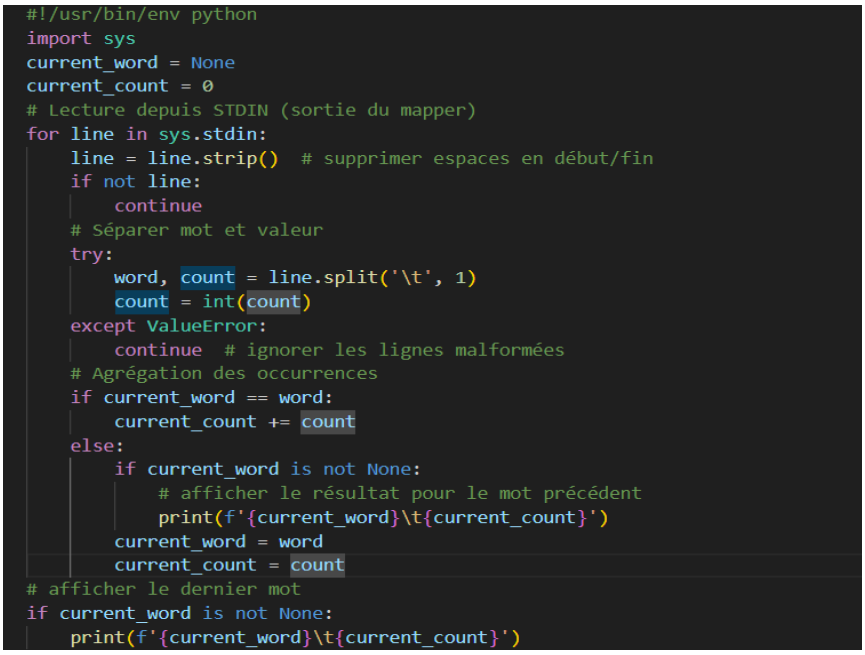
4)MapReduce avec python

Mapper.py

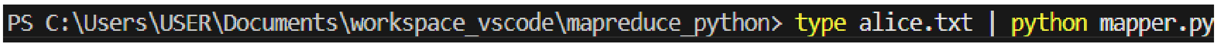


```
#!/usr/bin/env python
import sys

# input comes from standard input (STDIN)
for line in sys.stdin:
    line = line.strip()           # remove leading/trailing spaces
    words = line.split()          # split line into words
    for word in words:
        print('%s\t%s' % (word,1))       # output key-value pair to STDOUT
```

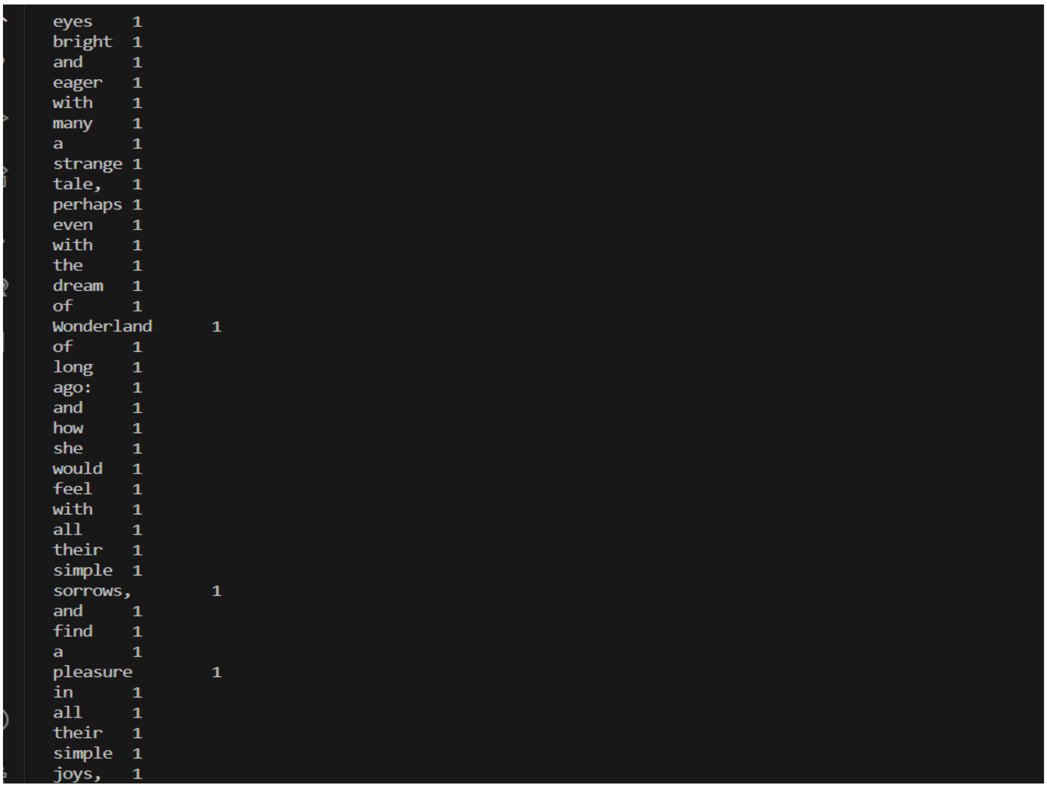
Reducer.py



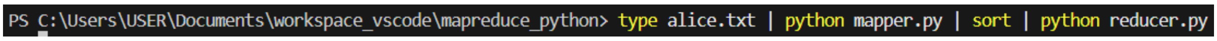
```
#!/usr/bin/env python
import sys
current_word = None
current_count = 0
# Lecture depuis STDIN (sortie du mapper)
for line in sys.stdin:
    line = line.strip()  # supprimer espaces en début/fin
    if not line:
        continue
    # Séparer mot et valeur
    try:
        word, count = line.split('\t', 1)
        count = int(count)
    except ValueError:
        continue  # ignorer les lignes malformées
    # Agrégation des occurrences
    if current_word == word:
        current_count += count
    else:
        if current_word is not None:
            # afficher le résultat pour le mot précédent
            print(f'{current_word}\t{current_count}')
        current_word = word
        current_count = count
# afficher le dernier mot
if current_word is not None:
    print(f'{current_word}\t{current_count}')
```



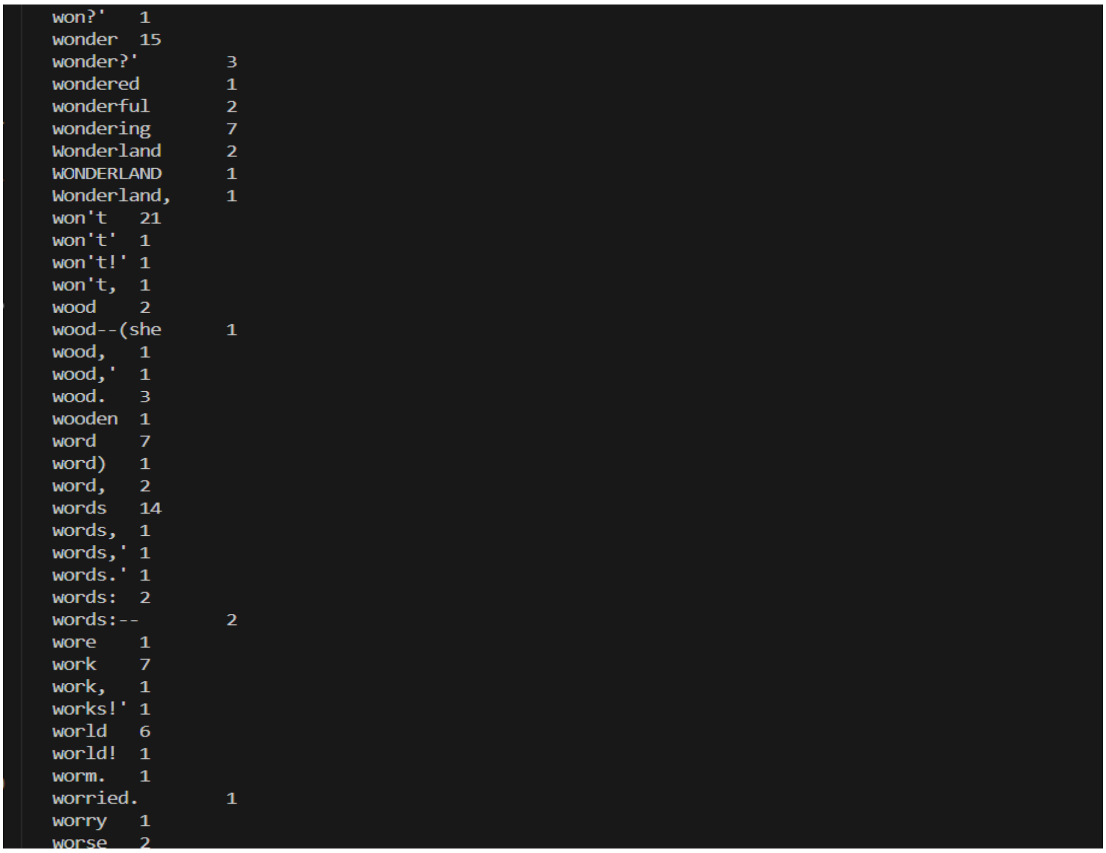
```
PS C:\Users\USER\Documents\workspace_vscode\mapreduce_python> type alice.txt | python mapper.py
```



```
eyes    1
bright  1
and     1
eager   1
with    1
many    1
a       1
strange 1
tale,   1
perhaps 1
even    1
with    1
the     1
dream   1
of      1
Wonderland      1
of      1
long    1
ago:    1
and     1
how     1
she     1
would   1
feel    1
with    1
all     1
their   1
simple  1
sorrows,        1
and     1
find    1
a       1
pleasure        1
in      1
all     1
their   1
simple  1
joys,   1
```

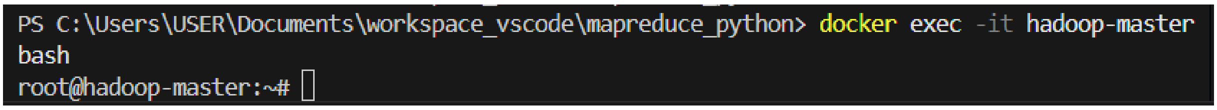


```
PS C:\Users\USER\Documents\workspace_vscode\mapreduce_python> type alice.txt | python mapper.py | sort | python reducer.py
```



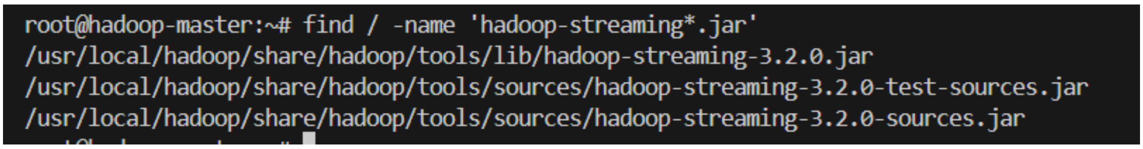
```
won?'   1
wonder  15
wonder?'        3
wondered        1
wonderful       2
wondering       7
Wonderland      2
WONDERLAND      1
Wonderland,     1
won't   21
won't'  1
won't!' 1
won't,  1
wood    2
wood--(she      1
wood,   1
wood,'  1
wood.   3
wooden  1
word    7
word)   1
word,   2
words   14
words,  1
words,' 1
words.' 1
words:  2
words:--        2
wore    1
work    7
work,   1
works!' 1
world   6
world!  1
worm.   1
worried.        1
worry   1
worse   2
```

ouvrir le terminal du container master

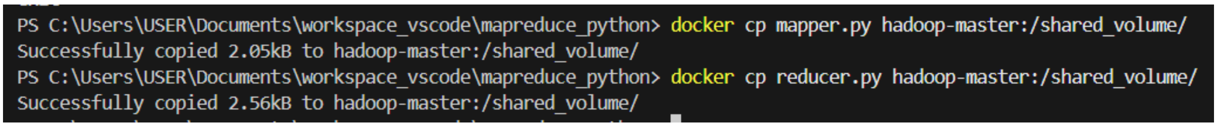


```
PS C:\Users\USER\Documents\workspace_vscode\mapreduce_python> docker exec -it hadoop-master
bash
root@hadoop-master:~# 
```

localiser le fichier JAR de l’utilitaire hadoop streaming

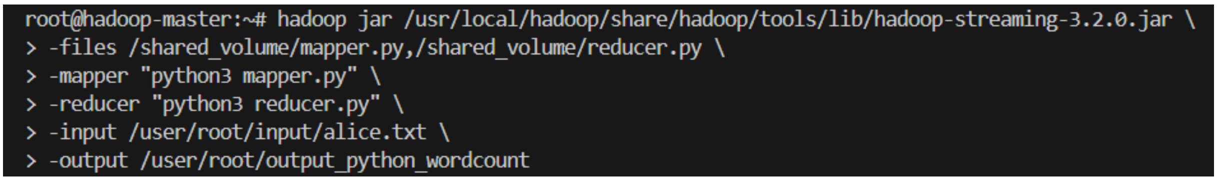


```
root@hadoop-master:~# find / -name 'hadoop-streaming*.jar'
/usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.2.0.jar
/usr/local/hadoop/share/hadoop/tools/sources/hadoop-streaming-3.2.0-test-sources.jar
/usr/local/hadoop/share/hadoop/tools/sources/hadoop-streaming-3.2.0-sources.jar
```

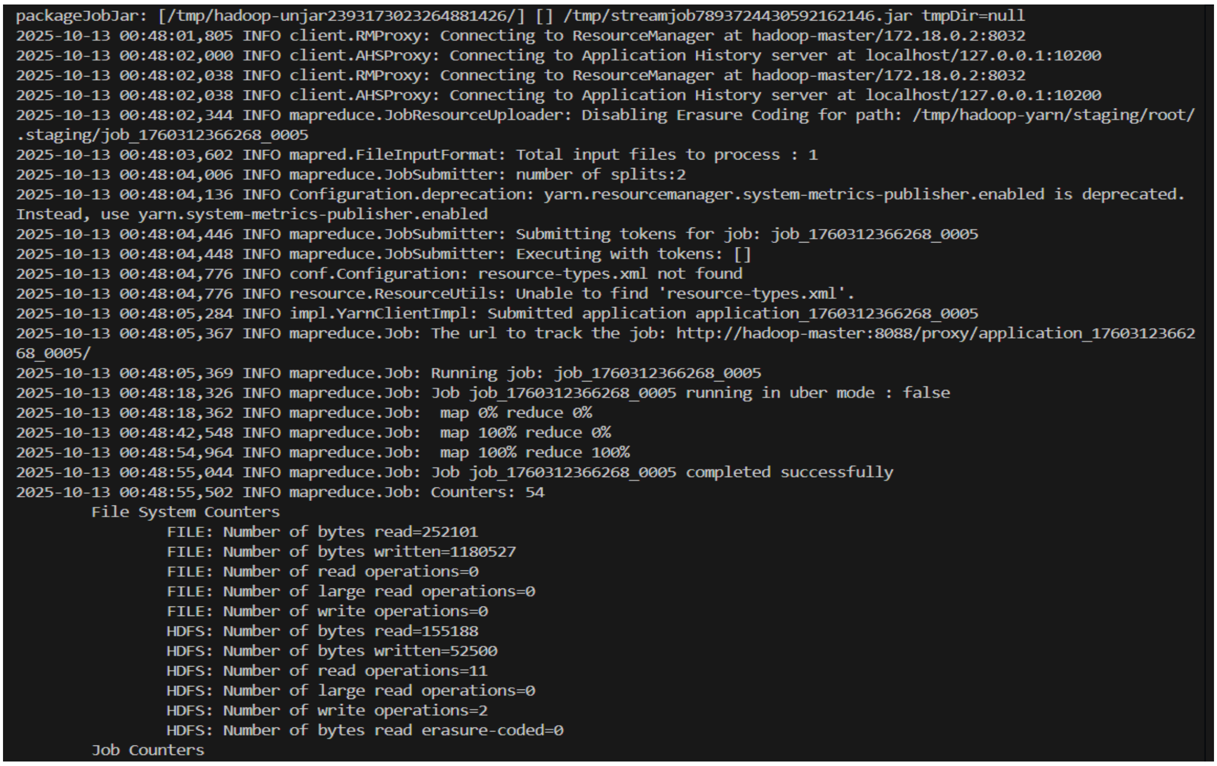


```
PS C:\Users\USER\Documents\workspace_vscode\mapreduce_python> docker cp mapper.py hadoop-master:/shared_volume/
Successfully copied 2.05kB to hadoop-master:/shared_volume/
PS C:\Users\USER\Documents\workspace_vscode\mapreduce_python> docker cp reducer.py hadoop-master:/shared_volume/
Successfully copied 2.56kB to hadoop-master:/shared_volume/
```

finalement exécuter le programme map/reduce

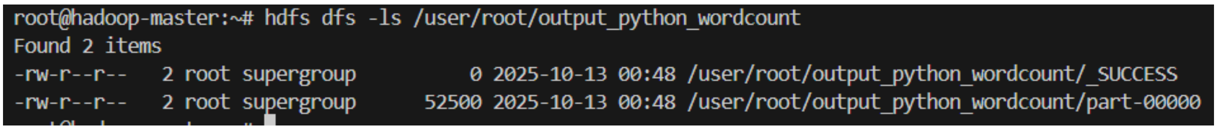


```
root@hadoop-master:~# hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.2.0.jar \
> -files /shared_volume/mapper.py,/shared_volume/reducer.py \
> -mapper "python3 mapper.py" \
> -reducer "python3 reducer.py" \
> -input /user/root/input/alice.txt \
> -output /user/root/output_python_wordcount
```



```
packageJobJar: [/tmp/hadoop-unjar2393173023264881426/] [] /tmp/streamjob7893724430592162146.jar tmpDir=null
2025-10-13 00:48:01,805 INFO client.RMProxy: Connecting to ResourceManager at hadoop-master/172.18.0.2:8032
2025-10-13 00:48:02,000 INFO client.AHSProxy: Connecting to Application History server at localhost/127.0.0.1:10200
2025-10-13 00:48:02,038 INFO client.RMProxy: Connecting to ResourceManager at hadoop-master/172.18.0.2:8032
2025-10-13 00:48:02,038 INFO client.AHSProxy: Connecting to Application History server at localhost/127.0.0.1:10200
2025-10-13 00:48:02,344 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/
.staging/job_1760312366268_0005
2025-10-13 00:48:03,602 INFO mapred.FileInputFormat: Total input files to process : 1
2025-10-13 00:48:04,006 INFO mapreduce.JobSubmitter: number of splits:2
2025-10-13 00:48:04,136 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated.
Instead, use yarn.system-metrics-publisher.enabled
2025-10-13 00:48:04,446 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1760312366268_0005
2025-10-13 00:48:04,448 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-10-13 00:48:04,776 INFO conf.Configuration: resource-types.xml not found
2025-10-13 00:48:04,776 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-10-13 00:48:05,284 INFO impl.YarnClientImpl: Submitted application application_1760312366268_0005
2025-10-13 00:48:05,367 INFO mapreduce.Job: The url to track the job: http://hadoop-master:8088/proxy/application_17603123662
68_0005/
2025-10-13 00:48:05,369 INFO mapreduce.Job: Running job: job_1760312366268_0005
2025-10-13 00:48:18,326 INFO mapreduce.Job: Job job_1760312366268_0005 running in uber mode : false
2025-10-13 00:48:18,362 INFO mapreduce.Job:  map 0% reduce 0%
2025-10-13 00:48:42,548 INFO mapreduce.Job:  map 100% reduce 0%
2025-10-13 00:48:54,964 INFO mapreduce.Job:  map 100% reduce 100%
2025-10-13 00:48:55,044 INFO mapreduce.Job: Job job_1760312366268_0005 completed successfully
2025-10-13 00:48:55,502 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=252101
                FILE: Number of bytes written=1180527
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=155188
                HDFS: Number of bytes written=52500
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
```

vérifier les résultats de l’exécution sur HDFS



```
root@hadoop-master:~# hdfs dfs -ls /user/root/output_python_wordcount
Found 2 items
-rw-r--r--   2 root supergroup          0 2025-10-13 00:48 /user/root/output_python_wordcount/_SUCCESS
-rw-r--r--   2 root supergroup      52500 2025-10-13 00:48 /user/root/output_python_wordcount/part-00000
```