

EXP NO: 2 RUN A BASIC WORD COUNT MAP REDUCE PROGRAM TO UNDERSTAND MAP REDUCE PARADIGM

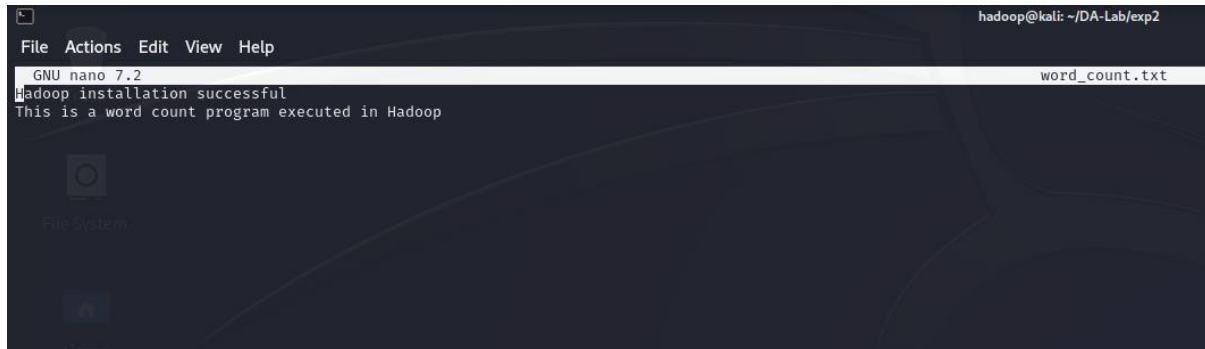
\$mkdir DA-Lab

\$cd DA-Lab

\$mkdir exp2

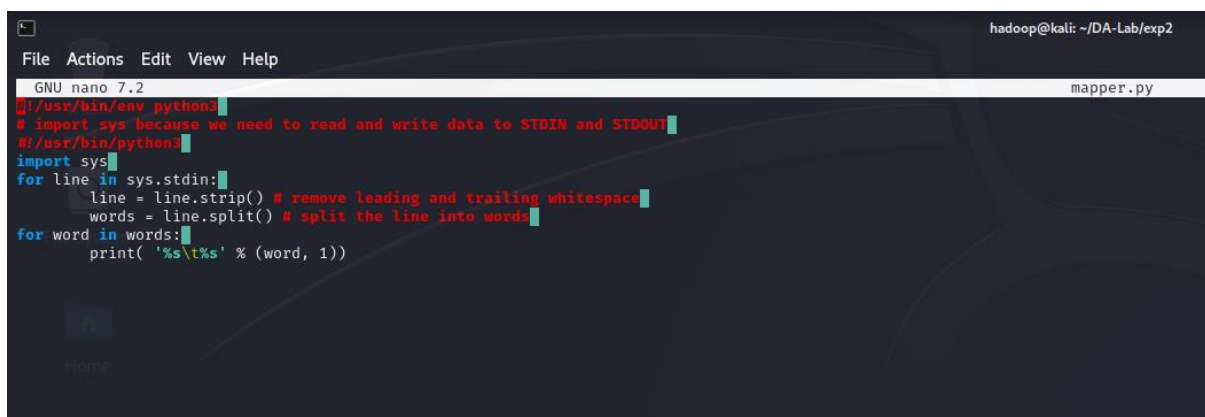
\$cd exp2

\$nano word_count.txt

A screenshot of a terminal window showing the nano text editor. The editor is open to a file named 'word_count.txt'. The text inside the file reads: 'Hadoop installation successful' followed by 'This is a word count program executed in Hadoop' on the next line. The terminal title bar shows 'hadoop@kali: ~/DA-Lab/exp2'.

```
hadoop@kali: ~/DA-Lab/exp2
GNU nano 7.2 word_count.txt
Hadoop installation successful
This is a word count program executed in Hadoop
```

\$nano mapper.py

A screenshot of a terminal window showing the nano text editor. The editor is open to a file named 'mapper.py'. The code is a Python script that reads from standard input, strips whitespace, splits the line into words, and prints each word followed by a tab and the number 1. The terminal title bar shows 'hadoop@kali: ~/DA-Lab/exp2'.

```
hadoop@kali: ~/DA-Lab/exp2
GNU nano 7.2 mapper.py
#!/usr/bin/env python3
# import sys because we need to read and write data to STDIN and STDOUT
#!/usr/bin/python3
import sys
for line in sys.stdin:
    line = line.strip() # remove leading and trailing whitespace
    words = line.split() # split the line into words
    for word in words:
        print( '%s\t%s' % (word, 1))
```

\$nano reducer.py

A screenshot of a terminal window showing the nano text editor. The editor is open to a file named 'reducer.py'. The code is a Python script that reads from standard input, splits the line by tab, increments a counter for the current word, and prints the word and its count when a new word is encountered or at the end of the input. The terminal title bar shows 'hadoop@kali: ~/DA-Lab/exp2'.

```
hadoop@kali: ~/DA-Lab/exp2
GNU nano 7.2 reducer.py
from operator import itemgetter
import sys
current_word = None
current_count = 0
word = None
for line in sys.stdin:
    line = line.strip()
    word, count = line.split('\t', 1)
    try:
        count = int(count)
    except ValueError:
        continue
    if current_word == word:
        current_count += count
    else:
        if current_word:
            print( '%s\t%s' % (current_word, current_count))
            current_count = count
            current_word = word
        if current_word == word:
            print( '%s\t%s' % (current_word, current_count))
```


\$hdfs dfs -cat /exp2/output/*

```
(hadoop@kali)-[~/hadoop/bin]
$ ./hdfs dfs -cat /exp2/output/*
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
2024-09-21 00:07:24,178 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
Hadoop 1
This 1
a 1
count 1
executed 1
in 1
is 1
program 1
word 1
```