**EXP 2**  210701201

**Run a basic Word Count Map Reduce program to understand Map Reduce Paradigm.**

**Aim:**

To run a basic Word Count MapReduce program.

**Procedure:**

**Step 1: Create Data File:**

Create a file named "word\_count\_data.txt" and populate it with text data that you wish to analyse.

Login with your hadoop user.

# nano word\_count.txt

Output: Type the below content in word\_count.txt

# Step 2: Mapper Logic - mapper.py:

Create a file named "mapper.py" to implement the logic for the mapper. The mapper will read input data from STDIN, split lines into words, and output each word with its count.



#!/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))

.

# Step 3: Reducer Logic - reducer.py:

Create a file named "reducer.py" to implement the logic for the reducer. The reducer will aggregate the occurrences of each word and generate the final output.



# reducer.py

#!/usr/bin/python3

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))

# Step 4: Prepare Hadoop Environment:

Start the Hadoop daemons and create a directory in HDFS to store your data.



# Step 6: Make Python Files Executable:

Give executable permissions to your mapper.py and reducer.py files.



# Step 7: Run Word Count using Hadoop Streaming:

Download the latest hadoop-streaming jar file and place it in a location you can easily

access.

Then run the Word Count program using Hadoop Streaming.

hadoop jar /path/to/hadoop-streaming-3.3.6.jar \

-input /word\_count\_in\_python/word\_count\_data.txt \

-output /word\_count\_in\_python/new\_output \

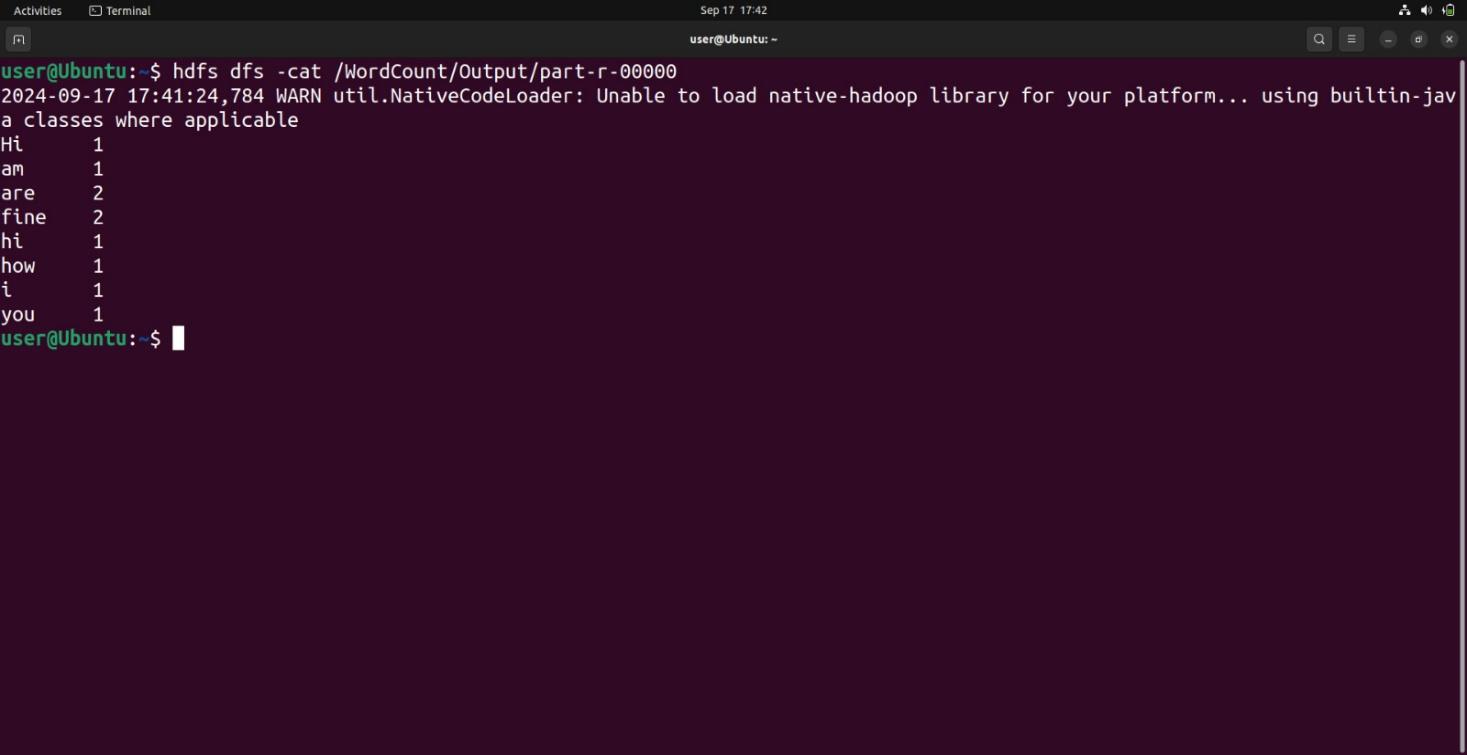
-mapper /path/to/mapper.py \

-reducer /path/to/reducer.py

# Step 8: Check Output:

Check the output of the Word Count program in the specified HDFS output directory

. hdfs dfs -cat /word\_count\_in\_python/new\_output/part-00000

**OUTPUT:**

**Result:**

Thus, the program for basic Word Count Map Reduce has been executed successfully.