**EXP 2** 210701249

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

**Aim:**

To run a basic Word Count MapReduce program.

**Procedure:**

**Step1:CreateDataFile:**

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

Loginwith yourhadoop user.

# nanoword\_count.txt

Output: Typethebelowcontent in word\_count.txt

# Step2:MapperLogic - mapper.py:

Create a file named "mapper.py" to implement the logic for the mapper. The mapperwillreadinputdatafrom STDIN,splitlinesintowords,andoutputeachwordwithitscount.

nanomapper.py

#Copyand pastethemapper.pycode

#!/usr/bin/envpython3

# import sys because we need to read and write data to STDIN and STDOUT#!/usr/bin/python3

importsys

forlinein sys.stdin:

line=line.strip() # remove leading and trailing whitespacewords= line.split()#split thelineinto words

forwordinwords:

print('%s\t%s'%(word,1))

.

# Step3:ReducerLogic-reducer.py:

Create a file named "reducer.py" to implement the logic for the reducer. The reducerwillaggregate theoccurrences of each word and generatethe final output.

nanoreducer.py

#Copyand pastethereducer.pycode

# reducer.py

#!/usr/bin/python3

from operator import itemgetterimport sys

current\_word = Nonecurrent\_count = 0word=None

for line in sys.stdin:line=line.strip()

word, count = line.split('\t', 1)try:

count = int(count)exceptValueError:

continue

if current\_word == word:current\_count+=count

else:

ifcurrent\_word:

print( '%s\t%s' % (current\_word, current\_count))current\_count= count

current\_word = wordifcurrent\_word==word:

print('%s\t%s'%(current\_word,current\_count))

# Step4:PrepareHadoopEnvironment:

StarttheHadoop daemonsand createadirectory inHDFS tostoreyourdata.

start-all.sh

hdfsdfs-mkdir/word\_count\_in\_python

hdfsdfs-copyFromLocal/path/to/word\_count.txt/word\_count\_in\_python

# Step6:MakePythonFilesExecutable:

Giveexecutablepermissionstoyourmapper.py andreducer.pyfiles.

chmod777mapper.pyreducer.py

# Step7:RunWordCountusingHadoopStreaming:

Downloadthelatest hadoop-streamingjarfileandplaceit inalocation youcaneasily

access.

ThenruntheWordCountprogramusingHadoopStreaming.

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

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

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

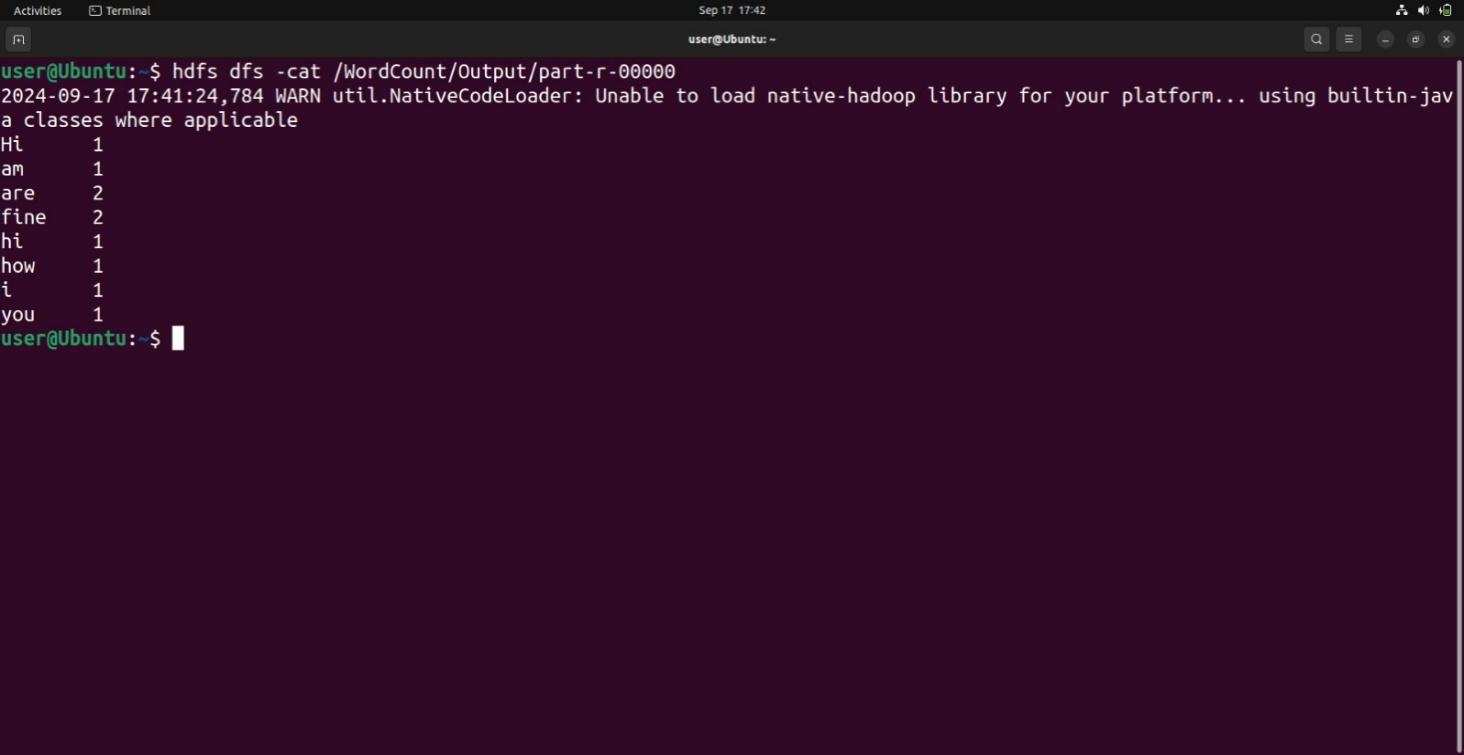
-mapper/path/to/mapper.py\

-reducer/path/to/reducer.py

# Step8:CheckOutput:

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

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

**OUTPUT:**

**Result:**

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