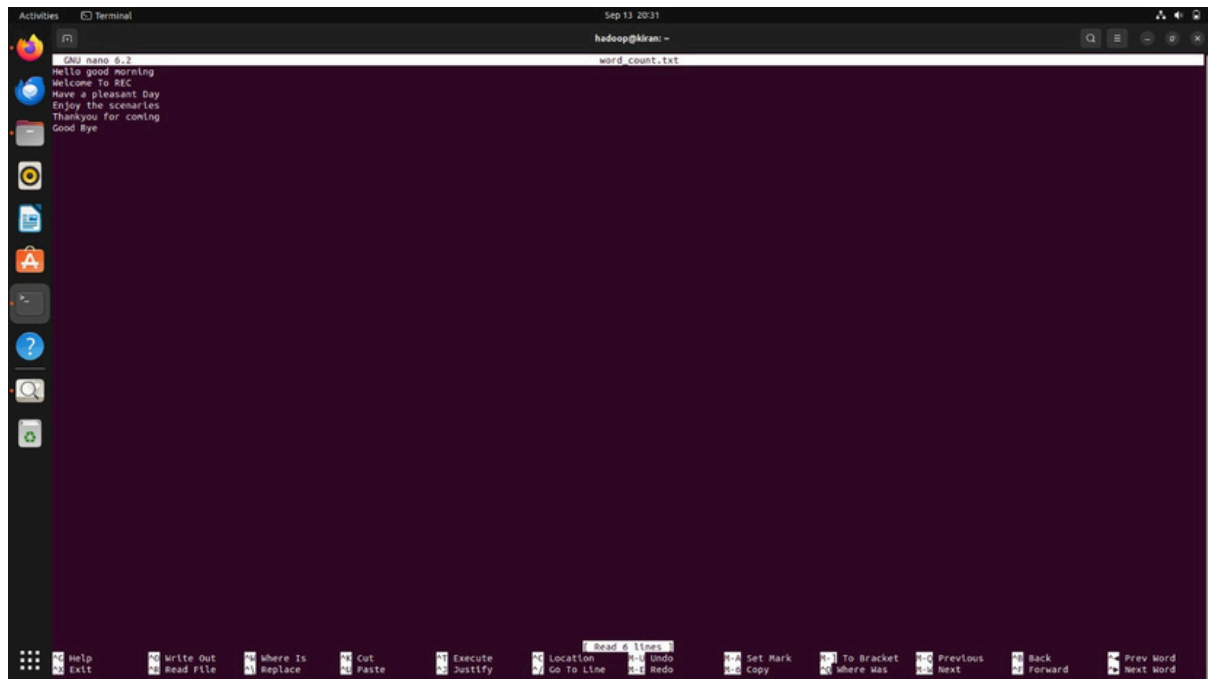


Exp. No : 2

Word Count Map Reduce program

1. Create word_count.txt file




The screenshot shows a terminal window with the nano editor open. The file being edited is named 'word_count.txt'. The content of the file is as follows:

```
GNU nano 6.2
Hello good morning
Welcome To REC
Have a pleasant Day
Enjoy the scenarios
Thankyou for coming
Good Bye
```

The terminal window title is 'hadoop@kiran: ~' and the date is 'Sep 13 2021'. The bottom status bar shows various nano editor commands like 'Help', 'Exit', 'Write Out', 'Read File', etc.

2. Create mapper.py program



The screenshot shows a terminal window with the nano editor open. The file being edited is named 'mapper.py'. The content of the file is as follows:

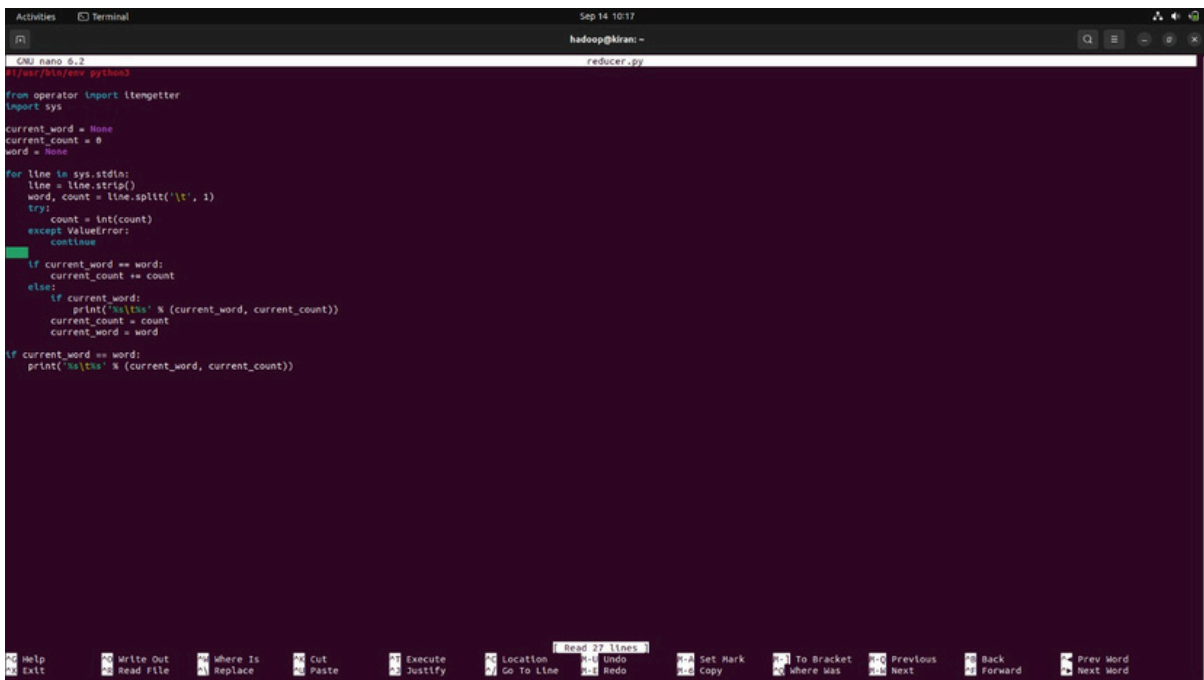
```
GNU nano 6.2
#!/usr/bin/env python3
import sys

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))

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))
```

The terminal window title is 'hadoop@kiran: ~' and the date is 'Sep 14 10:17'. The bottom status bar shows various nano editor commands like 'Help', 'Exit', 'Write Out', 'Read File', etc.

3. Create reducer.py program.



```

hadoop@kiran: ~
$ nano 6.2
$ cd /usr/local/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(' ', 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))

```

4. Storing the word_count.txt in HDFS Storage.

```

hadoop@koushik ~$ nano word_count.txt
hadoop@koushik ~$ hdfs dfs -cat /word_count_in_python/new_output/part-00000
Bye 1
Day 1
Enjoy 1
Good 1
Have 1
Hello 1
REC 1
Thankyou 1
To 1
Welcome 1
a 1
coming 1
for 1
good 1
morning 1
pleasant 1
scenarios 1
the 1

```

5. Running the Word Count program using Hadoop Streaming.

```

Activities Terminal Sep 14 10:18
hadoop@kiran: ~
2024-09-14 10:13:04.247 ERROR streaming.StreamJob: Error Launching Job : Output directory hdfs://localhost:9000/word_count_in_python/new_output already exists
Streaming Command Failed!
hadoop@kiran: ~
hadoop@kiran:~$ hdfs dfs -rm -r /word_count_in_python/new_output
deleted /word_count_in_python/new_output
hadoop@kiran:~$ hadoop jar /home/hadoop/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.6.jar \
-input /word_count_in_python/word_count.txt \
-output /word_count_in_python/new_output \
-mapper ./mapper.py \
-reducer ./reducer.py
2024-09-14 10:13:36.084 INFO Impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2024-09-14 10:13:36.082 INFO Impl.MetricsSystemImpl: Scheduled metric snapshot period at 10 second(s).
2024-09-14 10:13:36.082 INFO Impl.MetricsSystemImpl: JobTracker metrics system started
2024-09-14 10:13:36.021 WARN Impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2024-09-14 10:13:37.189 INFO mapred.FileInputFormat: Total input files to process : 1
2024-09-14 10:13:37.192 INFO mapreduce.JobSubmitter: number of splits:1
2024-09-14 10:13:37.333 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local682529798_0001
2024-09-14 10:13:37.333 INFO mapreduce.JobSubmitter: Executing with tokens: {}
2024-09-14 10:13:37.527 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2024-09-14 10:13:37.530 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
2024-09-14 10:13:37.534 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2024-09-14 10:13:37.536 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-09-14 10:13:37.538 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup_temporary folders under output directory:false, ignore cleanup failures: false
2024-09-14 10:13:37.537 INFO mapreduce.Job: Running Job: job_local682529798_0001
2024-09-14 10:13:37.602 INFO mapred.LocalJobRunner: Waiting for map tasks
2024-09-14 10:13:37.604 INFO mapred.LocalJobRunner: Starting task: attempt_local682529798_0001_r_000000_0
2024-09-14 10:13:37.644 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-09-14 10:13:37.645 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup_temporary folders under output directory:false, ignore cleanup failures: false
2024-09-14 10:13:37.679 INFO mapreduce.Task: Using ResourceCalculatorProcessFree: {}
2024-09-14 10:13:37.728 INFO mapred.MapTask: numReduceTasks: 1
2024-09-14 10:13:37.732 INFO mapred.MapTask: (EQUATOR) 0 kv 20214306104057504
2024-09-14 10:13:37.793 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2024-09-14 10:13:37.793 INFO mapred.MapTask: soft limit at 83886080
2024-09-14 10:13:37.793 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2024-09-14 10:13:37.793 INFO mapred.MapTask: kvstart = 20214306; length = 4553600
2024-09-14 10:13:37.796 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2024-09-14 10:13:37.799 INFO streaming.PipeMapred: PipeMapred exec [/home/hadoop/mapper.py]
2024-09-14 10:13:37.804 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir
2024-09-14 10:13:37.804 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.file
2024-09-14 10:13:37.805 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce.map.input.length
2024-09-14 10:13:37.805 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
2024-09-14 10:13:37.805 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
2024-09-14 10:13:37.806 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start
2024-09-14 10:13:37.807 INFO Configuration.deprecation: mapred.task.ismap is deprecated. Instead, use mapreduce.task.ismap
2024-09-14 10:13:37.807 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
2024-09-14 10:13:37.807 INFO Configuration.deprecation: mapred.tlp.id is deprecated. Instead, use mapreduce.task.id
2024-09-14 10:13:37.808 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
2024-09-14 10:13:37.808 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
2024-09-14 10:13:37.945 INFO streaming.PipeMapred: R/W=1/0/0 in:NA [rec/s] out:NA [rec/s]
2024-09-14 10:13:37.948 INFO streaming.PipeMapred: Records R/W=1/
2024-09-14 10:13:37.949 INFO streaming.PipeMapred: MErrorThread done
2024-09-14 10:13:37.950 INFO streaming.PipeMapred: mapred/Finished
2024-09-14 10:13:37.952 INFO mapred.LocalJobRunner:
2024-09-14 10:13:37.953 INFO mapred.MapTask: Starting flush of map output
2024-09-14 10:13:37.953 INFO mapred.MapTask: Spilling map output

```

```

Activities Terminal Sep 14 10:19
hadoop@kiran: ~
2024-09-14 10:13:38.349 INFO mapred.LocalJobRunner: Finishing task: attempt_local682529798_0001_r_000000_0
2024-09-14 10:13:38.350 INFO mapred.LocalJobRunner: reduce task executor complete.
2024-09-14 10:13:38.547 INFO mapreduce.Job: Job job_local682529798_0001 running in user mode : false
2024-09-14 10:13:38.549 INFO mapreduce.Job: map 100% reduce 100%
2024-09-14 10:13:38.552 INFO mapreduce.Job: Job job_local682529798_0001 completed successfully
2024-09-14 10:13:38.563 INFO mapreduce.Job: Counters: 36
File System Counters
  FILE: Number of bytes read=283202
  FILE: Number of bytes written=1572231
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=206
  HDFS: Number of bytes written=139
  HDFS: Number of read operations=15
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=4
  HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
  Map input records=0
  Map output records=18
  Map output bytes=139
  Map output materialized bytes=181
  Input split bytes=109
  Combine input records=0
  Combine output records=0
  Reduce input groups=18
  Reduce shuffle bytes=181
  Reduce input records=18
  Reduce output records=18
  Spilled Records=36
  Shuffled Maps=1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=16
  Total committed heap usage (bytes)=585413632
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=103
File Output Format Counters
  Bytes Written=139
2024-09-14 10:13:38.563 INFO streaming.StreamJob: Output directory: /word_count_in_python/new_output
hadoop@kiran:~$ hdfs dfs -cat /word_count_in_python/new_output/part-00000

```

Output :

```
hadoop@koushik ~$ nano word_count.txt
hadoop@koushik ~$ hdfs dfs -cat /word_count_in_python/new_output/part-00000
Bye      1
Day      1
Enjoy    1
Good     1
Have     1
Hello    1
REC      1
Thankyou      1
To        1
Welcome    1
a          1
coming     1
for        1
good       1
morning    1
pleasant   1
scenarios  1
the        1
```