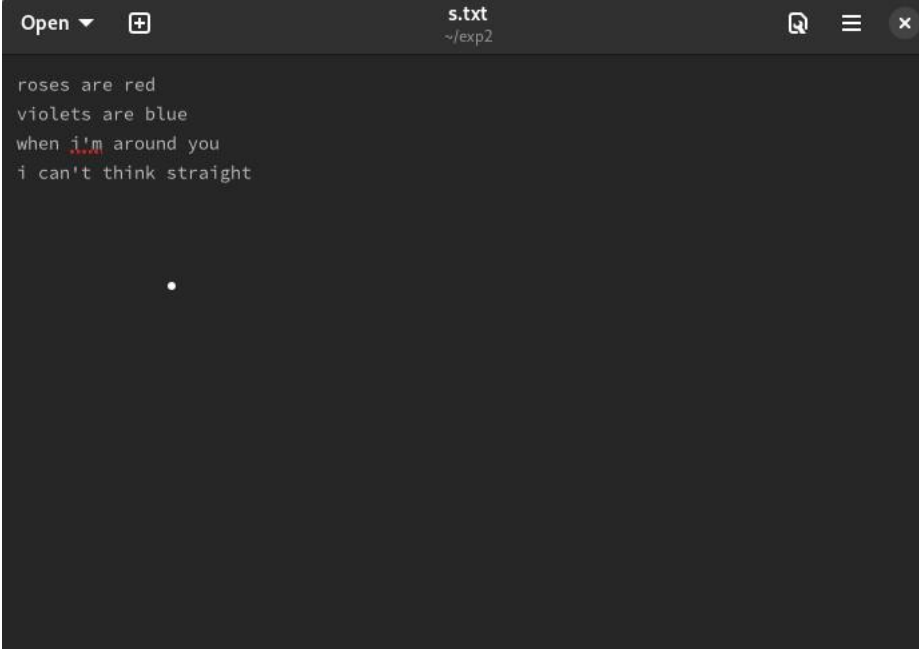



**Exp. No : 2****Word Count Map Reduce program****1. Create s.txt file**

A screenshot of a text editor window titled 's.txt' with a path of '~/.exp2'. The window contains the following text:

```
roses are red  
violets are blue  
when i'm around you  
i can't think straight
```

**2. Create mapper.py program**

A screenshot of a nano text editor window titled 'mapper.py'. The window shows the following Python code:

```
#!/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))
```

At the bottom of the window, there is a status bar with the text '[ Read 9 lines ]' and a table of keyboard shortcuts:

<b>^G</b> Help	<b>^O</b> Write Out	<b>^W</b> Where Is	<b>^K</b> Cut	<b>^T</b> Execute	<b>^C</b> Location
<b>^X</b> Exit	<b>^R</b> Read File	<b>^I</b> Replace	<b>^U</b> Paste	<b>^J</b> Justify	<b>^_</b> Go To Line

### 3. Create reducer.py program.

```

GNU nano 7.2                                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))

^G Help      ^O Write Out  ^W Where Is   ^K Cut        ^T Execute
^X Exit      ^R Read File  ^\ Replace    ^U Paste      ^J Justify

```

### 4. Running the Word Count program using Hadoop Streaming

```

keerthi@fedora:~$ echo $JAVA_HOME
/usr/lib/jvm/java-8-openjdk
keerthi@fedora:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as keerthi in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [fedora]
Starting resourcemanager
Starting nodemanagers
keerthi@fedora:~$ hadoop jar $HADOOP_STREAMING -input /exp1/s.txt -output /exp1/output1 -mapper ~/exp2/mapper.py -reducer ~/exp2/reducer.py
packageJobJar: [/tmp/hadoop-unjar6892827399228816367/] [] /tmp/streamjob5520682693925799301.jar tmpDir=null
2024-10-20 09:35:00,120 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-10-20 09:35:00,581 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-10-20 09:35:06,740 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/keerthi/.staging/job_1729431042099_0001
2024-10-20 09:35:08,047 INFO mapreduce.JobSubmitter: Cleaning up the staging area /tmp/hadoop-yarn/staging/keerthi/.staging/job_1729431042099_0001
2024-10-20 09:35:08,116 ERROR streaming.StreamJob: Error Launching job : Input path does not exist: hdfs://localhost:9000/exp1/s.txt
Streaming Command Failed!
keerthi@fedora:~$ hadoop jar $HADOOP_STREAMING -input /exp2/s.txt -output /exp2/output1 -mapper ~/exp2/mapper.py -reducer ~/exp2/reducer.py
packageJobJar: [/tmp/hadoop-unjar2409273401544148452/] [] /tmp/streamjob1526771308104542536.jar tmpDir=null
2024-10-20 09:44:16,211 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-10-20 09:44:16,530 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-10-20 09:44:17,201 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/keerthi/.staging/job_1729431042099_0002
2024-10-20 09:44:18,280 INFO mapred.FileInputFormat: Total input files to process : 1
2024-10-20 09:44:19,319 INFO mapreduce.JobSubmitter: number of splits:2
2024-10-20 09:44:20,068 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1729431042099_0002
2024-10-20 09:44:20,068 INFO mapreduce.JobSubmitter: Executing with tokens: []

```

```

2024-10-20 09:44:19,319 INFO mapreduce.JobSubmitter: number of splits:2
2024-10-20 09:44:20,068 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1729431042099_0002
2024-10-20 09:44:20,068 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-10-20 09:44:20,340 INFO conf.Configuration: resource-types.xml not found
2024-10-20 09:44:20,340 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-10-20 09:44:21,694 INFO impl.YarnClientImpl: Submitted application application_1729431042099_0002
2024-10-20 09:44:21,805 INFO mapreduce.Job: The url to track the job: http://fedora:8088/proxy/application_1729431042099_0002/
2024-10-20 09:44:21,808 INFO mapreduce.Job: Running job: job_1729431042099_0002
2024-10-20 09:44:42,322 INFO mapreduce.Job: Job job_1729431042099_0002 running in uber mode : false
2024-10-20 09:44:42,325 INFO mapreduce.Job: map 0% reduce 0%
2024-10-20 09:45:02,749 INFO mapreduce.Job: map 50% reduce 0%
2024-10-20 09:45:03,759 INFO mapreduce.Job: map 100% reduce 0%
2024-10-20 09:45:15,885 INFO mapreduce.Job: map 100% reduce 100%
2024-10-20 09:45:18,988 INFO mapreduce.Job: Job job_1729431042099_0002 completed successfully
2024-10-20 09:45:19,157 INFO mapreduce.Job: Counters: 54
    File System Counters
        FILE: Number of bytes read=278
        FILE: Number of bytes written=835515
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=396
        HDFS: Number of bytes written=175
        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
        Launched map tasks=2
        Launched reduce tasks=1
        Data-local map tasks=2
        Total time spent by all maps in occupied slots (ms)=14272
        Total time spent by all reduces in occupied slots (ms)=9946

```

```

keerthi@fedora:~ — /usr/lib/jvm/java-8-openjdk/bin/java -Dproc_jar -Djava.library.path=/home/keerthi/hadoop/lib/native -Dyarn.L...
Launched map tasks=2
Launched reduce tasks=1
Data-local map tasks=2
Total time spent by all maps in occupied slots (ms)=14272
Total time spent by all reduces in occupied slots (ms)=9946
Total time spent by all map tasks (ms)=14272
Total time spent by all reduce tasks (ms)=9946
Total vcore-milliseconds taken by all map tasks=14272
Total vcore-milliseconds taken by all reduce tasks=9946
Total megabyte-milliseconds taken by all map tasks=14614528
Total megabyte-milliseconds taken by all reduce tasks=10184704
Map-Reduce Framework
    Map input records=7
    Map output records=30
    Map output bytes=212
    Map output materialized bytes=284
    Input split bytes=168
    Combine input records=0
    Combine output records=0
    Reduce input groups=24
    Reduce shuffle bytes=284
    Reduce input records=30
    Reduce output records=24
    Spilled Records=60
    Shuffled Maps =2
    Failed Shuffles=0
    Merged Map outputs=2
    GC time elapsed (ms)=321
    CPU time spent (ms)=5060
    Physical memory (bytes) snapshot=878067712
    Virtual memory (bytes) snapshot=7738617856
    Total committed heap usage (bytes)=694681600
    Peak Map Physical memory (bytes)=320118784

```

```

keerthi@fedora:~ -- /usr/lib/jvm/java-8-openjdk/bin/java -Dproc_jar -Djava.library.path=/home/keerthi/hadoop/lib/native -Dyarn.L...
Input split bytes=168
Combine input records=0
Combine output records=0
Reduce input groups=24
Reduce shuffle bytes=284
Reduce input records=30
Reduce output records=24
Spilled Records=60
Shuffled Maps =2
Failed Shuffles=0
Merged Map outputs=2
GC time elapsed (ms)=321
CPU time spent (ms)=5060
Physical memory (bytes) snapshot=878067712
Virtual memory (bytes) snapshot=7738617856
Total committed heap usage (bytes)=694681600
Peak Map Physical memory (bytes)=320118784
Peak Map Virtual memory (bytes)=2576637952
Peak Reduce Physical memory (bytes)=240435200
Peak Reduce Virtual memory (bytes)=2585477120
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=228
File Output Format Counters
Bytes Written=175
024-10-20 09:45:19,158 INFO streaming.StreamJob: Output directory: /exp2/output1
keerthi@fedora:~$

```

```

keerthi@fedora:~$ hdfs dfs -cat/exp1/output1/part-00000
-cat/exp1/output1/part-00000: Unknown command
Usage: hadoop fs [generic options]
[-appendToFile [-n] <localsrc> ... <dst>]
[-cat [-ignoreCrc] <src> ...]
[-checksum [-v] <src> ...]
[-chgrp [-R] GROUP PATH...]
[-chmod [-R] <MODE[,MODE]... | OCTALMODE> PATH...]
[-chown [-R] [OWNER][:[GROUP]] PATH...]
[-concat <target path> <src path> <src path> ...]
[-copyFromLocal [-f] [-p] [-l] [-d] [-t <thread count>] [-q <thread pool queue size>] <localsrc> ... <dst>]
[-copyToLocal [-f] [-p] [-crc] [-ignoreCrc] [-t <thread count>] [-q <thread pool queue size>] <src> ... <localdst>]
[-count [-q] [-h] [-v] [-t <storage type>]] [-u] [-x] [-e] [-s] <path> ...]
[-cp [-f] [-p | -p[topax]] [-d] [-t <thread count>] [-q <thread pool queue size>] <src> ... <dst>]
[-createSnapshot <snapshotDir> [<snapshotName>]]
[-deleteSnapshot <snapshotDir> <snapshotName>]
[-df [-h] [<path> ...]]
[-du [-s] [-h] [-v] [-x] <path> ...]
[-expunge [-immediate] [-fs <path>]]
[-find <path> ... <expression> ...]
[-get [-f] [-p] [-crc] [-ignoreCrc] [-t <thread count>] [-q <thread pool queue size>] <src> ... <localdst>]
[-getfacl [-R] <path>]
[-getfattr [-R] {-n name | -d} [-e en] <path>]
[-getmerge [-nl] [-skip-empty-file] <src> <localdst>]
[-head <file>]
[-help [cmd ...]]
[-ls [-C] [-d] [-h] [-q] [-R] [-t] [-S] [-r] [-u] [-e] [<path> ...]]
[-mkdir [-p] <path> ...]
[-moveFromLocal [-f] [-p] [-l] [-d] <localsrc> ... <dst>]
[-moveToLocal <src> <localdst>]
[-mv <src> ... <dst>]
[-put [-f] [-p] [-l] [-d] [-t <thread count>] [-q <thread pool queue size>] <localsrc> ... <dst>]
[-renameSnapshot <snapshotDir> <oldName> <newName>]

```

**Output :**

```
keerthi@fedora:~$ hdfs dfs -cat /exp2/output/part-00000
Callin 1
Finally 1
LA 2
Lookin 1
Lost 1
Made 1
Maria 2
Might 1
Trynnna 1
dive 1
dough 1
for 2
in 2
it 1
make 1
marina 1
my 1
own 1
the 2
though 1
to 1
weed 1
without 1
yeah 2
```