3.Design a distributed application using MapReduce under Hadoop for: a) Character counting in a given text file. b) Counting no. of occurrences of every word in a given text file.

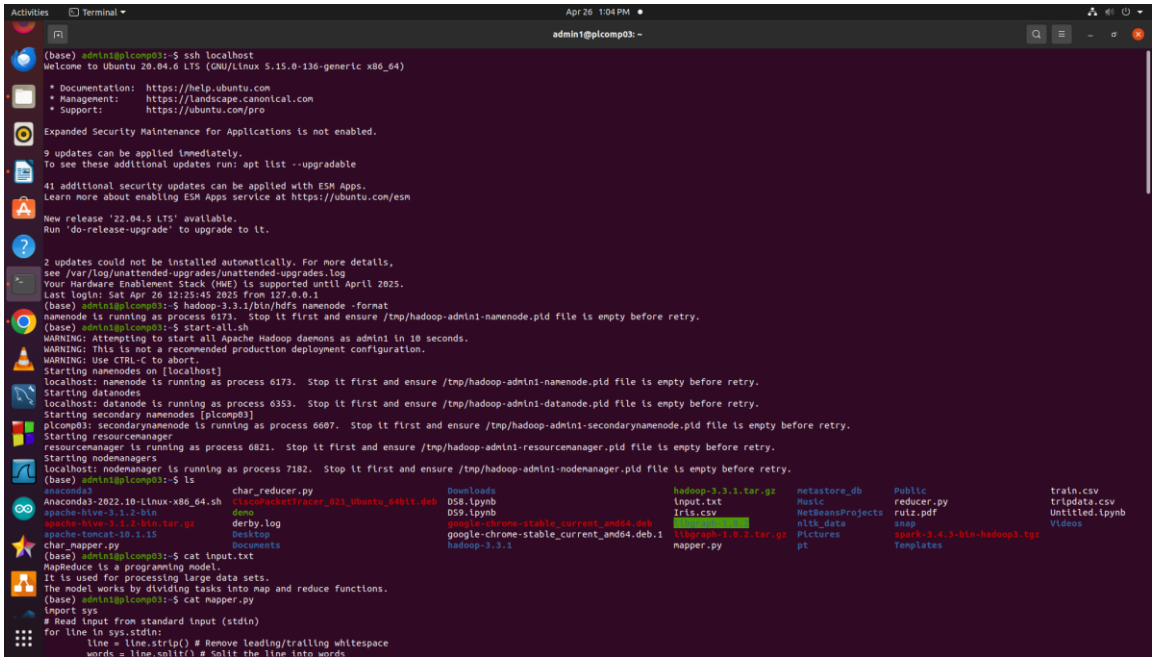
a)Character counting in a given text file:

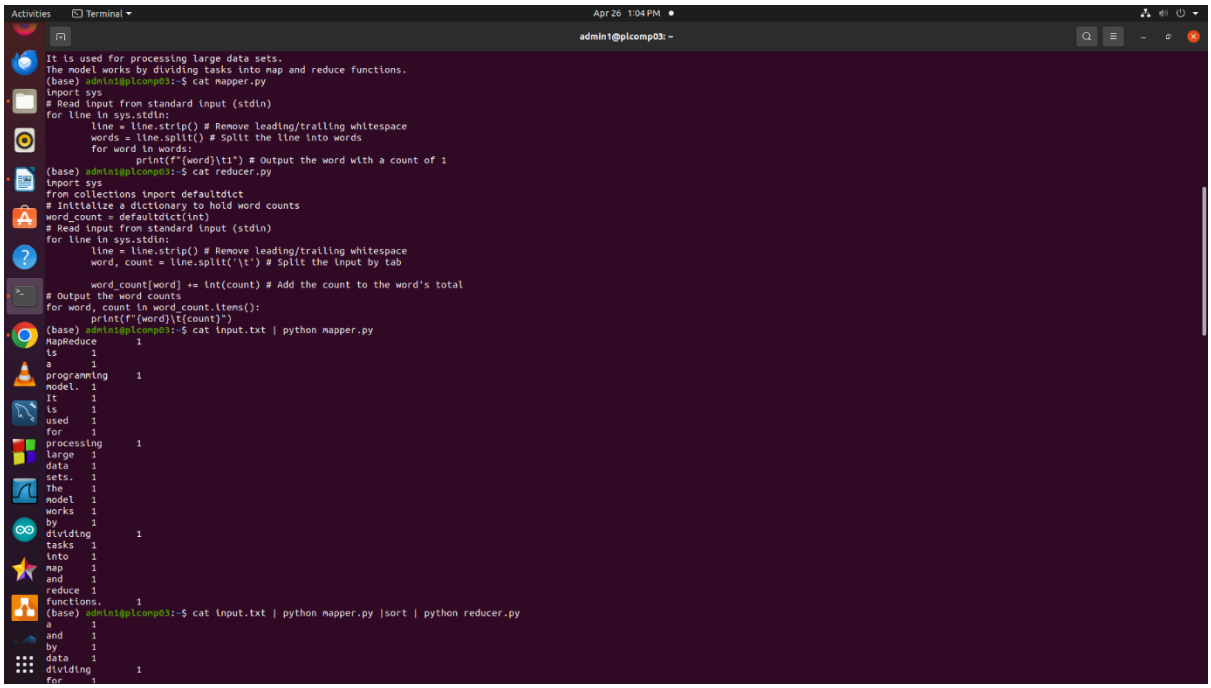
char_mapper.py

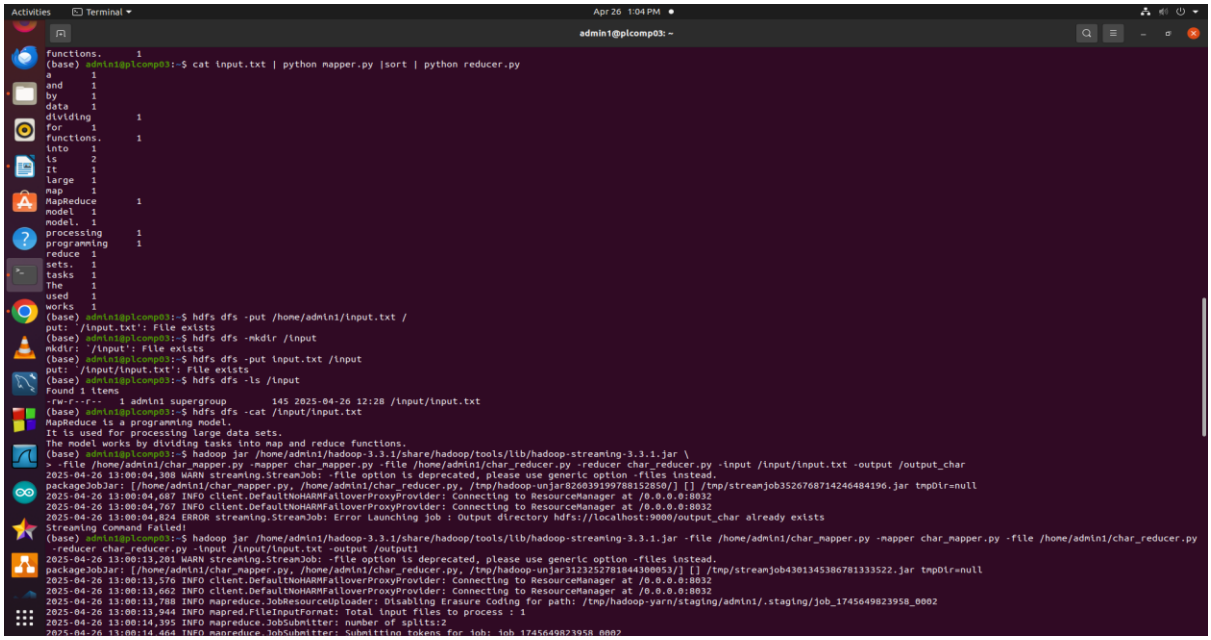
```
#!/usr/bin/env python3

import sys

# Read input line by line

for line in sys.stdin:

    line = line.strip()

    for char in line:

        if char != ' ':  # Ignore spaces if not needed

            print(f"{char}\t1")
```

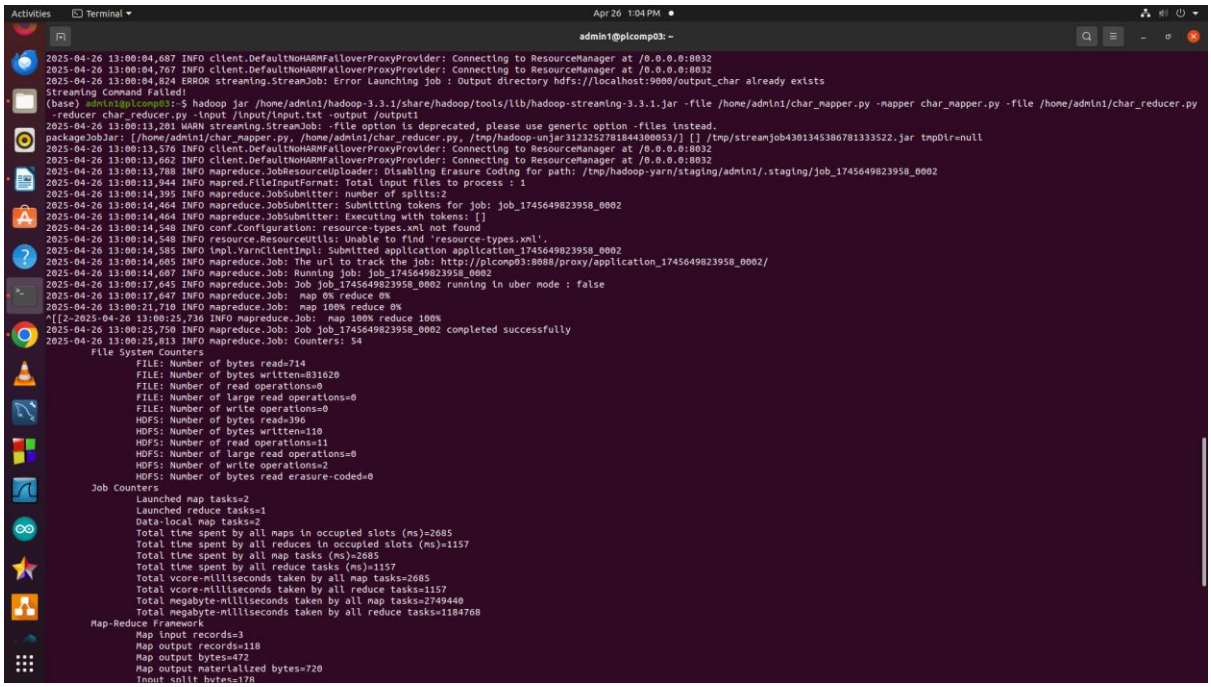
char_reducer.py

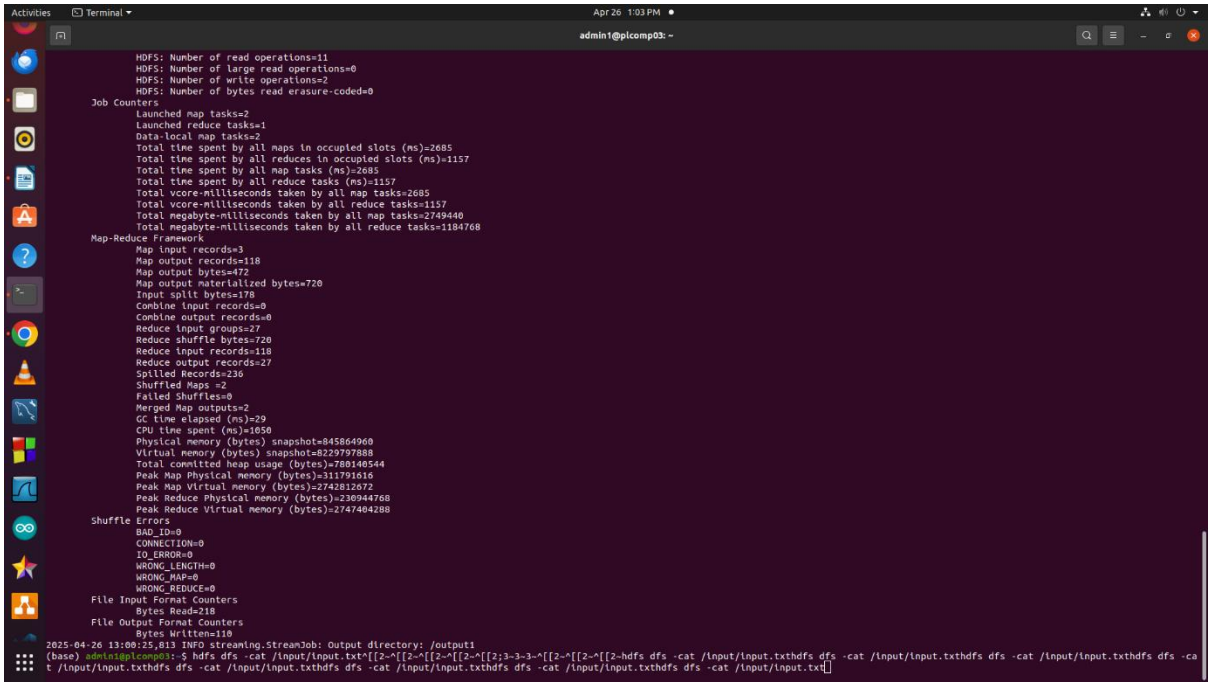
```
#!/usr/bin/env python3
import sys
from collections import defaultdict

char_count = defaultdict(int)

# Process input lines
for line in sys.stdin:
    line = line.strip()
    char, count = line.split('\t')
    char_count[char] += int(count)

# Output the character counts
for char, count in char_count.items():
    print(f"{char}\t{count}")
```

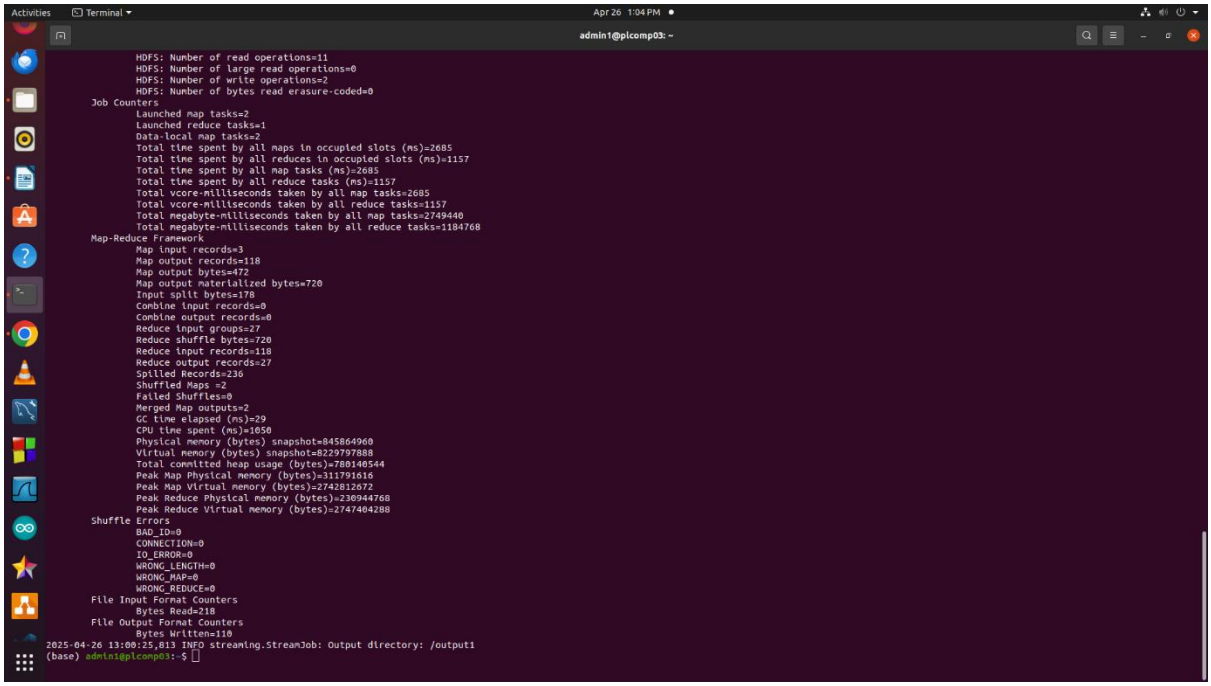

Activities
Terminal
Apr 26 1:04 PM
admin1@plcomp03: ~
(base) admin1@plcomp03:~$ ssh localhost
Welcome to Ubuntu 20.04.6 LTS (GNU/Linux 5.15.0-136-generic x86_64)
* Documentation: https://help.ubuntu.com
* Management: https://landscape.canonical.com
* Support: https://ubuntu.com/pro
Expanded Security Maintenance for Applications is not enabled.
9 updates can be applied immediately.
To see these additional updates run: apt list --upgradable
41 additional security updates can be applied with ESM Apps.
Learn more about enabling ESM Apps service at https://ubuntu.com/esm
New release '22.04.5 LTS' available.
Run 'do-release-upgrade' to upgrade to it.
2 updates could not be installed automatically. For more details,
see /var/log/unattended-upgrades/unattended-upgrades.log
Your Hardware Enablement Stack (HWE) is supported until April 2025.
Last login: Sat Apr 26 12:25:45 2025 from 127.0.0.1
(base) admin1@plcomp03:~$ hadoop-3.3.1/bin/hdfs namenode -format
namenode is running as process 6173. Stop it first and ensure /tmp/hadoop-admin1-namenode.pid file is empty before retry.
(base) admin1@plcomp03:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as admin1 in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
localhost: namenode is running as process 6173. Stop it first and ensure /tmp/hadoop-admin1-namenode.pid file is empty before retry.
Starting datanodes
localhost: datanode is running as process 6353. Stop it first and ensure /tmp/hadoop-admin1-datanode.pid file is empty before retry.
Starting secondary namenodes [plcomp03]
plcomp03: secondarynamenode is running as process 6607. Stop it first and ensure /tmp/hadoop-admin1-secondarynamenode.pid file is empty before retry.
Starting resourcemanager
resourcemanager is running as process 6821. Stop it first and ensure /tmp/hadoop-admin1-resourcemanager.pid file is empty before retry.
Starting nodemanagers
localhost: nodemanager is running as process 7182. Stop it first and ensure /tmp/hadoop-admin1-nodemanager.pid file is empty before retry.
(base) admin1@plcomp03:~$ ls
anaconda3
Anaconda3-2022.10-Linux-x86_64.sh
apache-hive-3.1.2-bin
apache-hive-3.1.2-bin.tar.gz
apache-tomcat-10.1.15
char_mapper.py
char_reducer.py
CiscoPacketTracer_821_Ubuntu_64bit.deb
demo
derby.log
Desktop
Documents
Downloads
DS8.ipynb
DS9.ipynb
google-chrome-stable_current_amd64.deb
google-chrome-stable_current_amd64.deb.1
hadoop-3.3.1
hadoop-3.3.1.tar.gz
input.txt
Iris.csv
libgraph-1.0.2.tar.gz
mapper.py
metastore_db
Music
NetBeansProjects
nltk_data
Pictures
pt
Public
reducer.py
ruiz.pdf
snap
Templates
train.csv
tripdata.csv
Untitled.ipynb
Videos
spark-3.4.3-bin-hadoop3.tgz
(base) admin1@plcomp03:~$ cat input.txt
MapReduce is a programming model.
It is used for processing large data sets.
The model works by dividing tasks into map and reduce functions.
(base) admin1@plcomp03:~$ cat mapper.py
import sys
# Read input from standard input (stdin)
for line in sys.stdin:
line = line.strip() # Remove leading/trailing whitespace
words = line.split() # Split the line into words


Activities
Terminal
Apr 26 1:04 PM
admin1@plcomp03: ~
It is used for processing large data sets.
The model works by dividing tasks into map and reduce functions.
(base) admin1@plcomp03:~$ cat mapper.py
import sys
# Read input from standard input (stdin)
for line in sys.stdin:
line = line.strip() # Remove leading/trailing whitespace
words = line.split() # Split the line into words
for word in words:
print(f"{word}\t1") # Output the word with a count of 1
(base) admin1@plcomp03:~$ cat reducer.py
import sys
from collections import defaultdict
# Initialize a dictionary to hold word counts
word_count = defaultdict(int)
# Read input from standard input (stdin)
for line in sys.stdin:
line = line.strip() # Remove leading/trailing whitespace
word, count = line.split('\t') # Split the input by tab
word_count[word] += int(count) # Add the count to the word's total
# Output the word counts
for word, count in word_count.items():
print(f"{word}\t{count}")
(base) admin1@plcomp03:~$ cat input.txt | python mapper.py
MapReduce 1
is 1
a 1
programming 1
model. 1
It 1
is 1
used 1
for 1
processing 1
large 1
data 1
sets. 1
The 1
model 1
works 1
by 1
dividing 1
tasks 1
into 1
map 1
and 1
reduce 1
functions. 1
(base) admin1@plcomp03:~$ cat input.txt | python mapper.py |sort | python reducer.py
a 1
and 1
by 1
data 1
dividing 1
for 1


Activities
Terminal
Apr 26 1:04 PM
admin1@plcomp03: ~
functions. 1
(base) admin1@plcomp03:~$ cat input.txt | python mapper.py |sort | python reducer.py
a 1
and 1
by 1
data 1
dividing 1
for 1
functions. 1
into 1
is 2
It 1
large 1
map 1
MapReduce 1
model 1
model. 1
processing 1
programming 1
reduce 1
sets. 1
tasks 1
The 1
used 1
works 1
(base) admin1@plcomp03:~$ hdfs dfs -put /home/admin1/input.txt /
put: `/input.txt': File exists
(base) admin1@plcomp03:~$ hdfs dfs -mkdir /input
mkdir: `/input': File exists
(base) admin1@plcomp03:~$ hdfs dfs -put input.txt /input
put: `/input/input.txt': File exists
(base) admin1@plcomp03:~$ hdfs dfs -ls /input
Found 1 items
-rw-r--r-- 1 admin1 supergroup 145 2025-04-26 12:28 /input/input.txt
(base) admin1@plcomp03:~$ hdfs dfs -cat /input/input.txt
MapReduce is a programming model.
It is used for processing large data sets.
The model works by dividing tasks into map and reduce functions.
(base) admin1@plcomp03:~$ hadoop jar /home/admin1/hadoop-3.3.1/share/hadoop/tools/lib/hadoop-streaming-3.3.1.jar \
> -file /home/admin1/char_mapper.py -mapper char_mapper.py -file /home/admin1/char_reducer.py -reducer char_reducer.py -input /input/input.txt -output /output_char
2025-04-26 13:00:04,308 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/home/admin1/char_mapper.py, /home/admin1/char_reducer.py, /tmp/hadoop-unjar8260391997881528 50/] [] /tmp/streamjob3526768714246484196.jar tmpDir=null
2025-04-26 13:00:04,607 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2025-04-26 13:00:04,767 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2025-04-26 13:00:04,824 ERROR streaming.StreamJob: Error Launching job : Output directory hdfs://localhost:9000/output_char already exists
Streaming Command Failed!
(base) admin1@plcomp03:~$ hadoop jar /home/admin1/hadoop-3.3.1/share/hadoop/tools/lib/hadoop-streaming-3.3.1.jar -file /home/admin1/char_mapper.py -mapper char_mapper.py -file /home/admin1/char_reducer.py
-reducer char_reducer.py -input /input/input.txt -output /output1
2025-04-26 13:00:13,201 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/home/admin1/char_mapper.py, /home/admin1/char_reducer.py, /tmp/hadoop-unjar3123252781844300053/] [] /tmp/streamjob4301345386781333522.jar tmpDir=null
2025-04-26 13:00:13,576 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2025-04-26 13:00:13,662 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2025-04-26 13:00:13,788 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/admin1/.staging/job_1745649823958_0002
2025-04-26 13:00:13,944 INFO mapred.FileInputFormat: Total input files to process : 1
2025-04-26 13:00:14,395 INFO mapreduce.JobSubmitter: number of splits:2
2025-04-26 13:00:14,464 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1745649823958_0002


Activities
Terminal
Apr 26 1:04 PM
admin1@plcomp03: ~
2025-04-26 13:00:04,607 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2025-04-26 13:00:04,767 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2025-04-26 13:00:04,824 ERROR streaming.StreamJob: Error Launching job : Output directory hdfs://localhost:9000/output_char already exists
Streaming Command Failed!
(base) admin1@plcomp03:~$ hadoop jar /home/admin1/hadoop-3.3.1/share/hadoop/tools/lib/hadoop-streaming-3.3.1.jar -file /home/admin1/char_mapper.py -mapper char_mapper.py -file /home/admin1/char_reducer.py
-reducer char_reducer.py -input /input/input.txt -output /output1
2025-04-26 13:00:13,201 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/home/admin1/char_mapper.py, /home/admin1/char_reducer.py, /tmp/hadoop-unjar3123252781844300053/] [] /tmp/streamjob4301345386781333522.jar tmpDir=null
2025-04-26 13:00:13,576 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2025-04-26 13:00:13,662 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2025-04-26 13:00:13,788 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/admin1/.staging/job_1745649823958_0002
2025-04-26 13:00:13,944 INFO mapred.FileInputFormat: Total input files to process : 1
2025-04-26 13:00:14,395 INFO mapreduce.JobSubmitter: number of splits:2
2025-04-26 13:00:14,464 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1745649823958_0002
2025-04-26 13:00:14,464 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-04-26 13:00:14,548 INFO conf.Configuration: resource-types.xml not found
2025-04-26 13:00:14,548 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-04-26 13:00:14,585 INFO impl.YarnClientImpl: Submitted application application_1745649823958_0002
2025-04-26 13:00:14,605 INFO mapreduce.Job: The url to track the job: http://plcomp03:8088/proxy/application_1745649823958_0002/
2025-04-26 13:00:14,607 INFO mapreduce.Job: Running job: job_1745649823958_0002
2025-04-26 13:00:17,645 INFO mapreduce.Job: Job job_1745649823958_0002 running in uber mode : false
2025-04-26 13:00:17,647 INFO mapreduce.Job: map 0% reduce 0%
2025-04-26 13:00:21,710 INFO mapreduce.Job: map 100% reduce 0%
^[[2~2025-04-26 13:00:25,736 INFO mapreduce.Job: map 100% reduce 100%
2025-04-26 13:00:25,750 INFO mapreduce.Job: Job job_1745649823958_0002 completed successfully
2025-04-26 13:00:25,813 INFO mapreduce.Job: Counters: 54
File System Counters
FILE: Number of bytes read=714
FILE: Number of bytes written=831620
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=110
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)=2685
Total time spent by all reduces in occupied slots (ms)=1157
Total time spent by all map tasks (ms)=2685
Total time spent by all reduce tasks (ms)=1157
Total vcore-milliseconds taken by all map tasks=2685
Total vcore-milliseconds taken by all reduce tasks=1157
Total megabyte-milliseconds taken by all map tasks=2749440
Total megabyte-milliseconds taken by all reduce tasks=1184768
Map-Reduce Framework
Map input records=3
Map output records=118
Map output bytes=472
Map output materialized bytes=720
Input split bytes=178





For Starting Hadoop any time use Following Commands on Terminal

1) ssh localhost

2) hadoop-3.3.1/bin/hdfs namenode -format

3) start-all.sh

4) Open the Browser and Type:- localhost:9870

check files

ls

check contents

cat input.txt

cat char_mapper.py

car char_reducer.py

For simple output

cat input.txt | python char_mapper.py

cat input.txt | python char_mapper.py |sort | python char_reducer.py

Access HDFS

hdfs dfs -put /home/admin1/input.txt /

hdfs dfs -mkdir /input

hdfs dfs -put input.txt /input

hdfs dfs -ls /input

hdfs dfs -cat /input/input.txt

Run the program:

hadoop jar /home/admin1/hadoop-3.3.1/share/hadoop/tools/lib/hadoop-streaming-3.3.1.jar \
-file /home/admin1/char_mapper.py -mapper char_mapper.py -file /home/admin1/char_reducer.py -reducer char_reducer.py -input /input/input.txt -output /output_char

See the output:

hdfs dfs -ls /output

hdfs dfs -cat /output/part-00000

b) Counting no. of occurrences of every word in a given text file.

mapper.py

```
import sys


# Read input from standard input (stdin)
for line in sys.stdin:
        line = line.strip()        # Remove leading/trailing whitespace
        words = line.split()   # Split the line into words
        for word in words:
                print(f"{word}\t1")      # Output the word with a count of 1
```

reducer.py

```
import sys
from collections import defaultdict


# Initialize a dictionary to hold word counts
word_count = defaultdict(int)


# Read input from standard input (stdin)
for line in sys.stdin:
        line = line.strip() # Remove leading/trailing whitespace
        word, count = line.split('\t') # Split the input by tab

        word_count[word] += int(count) # Add the count to the word's total


# Output the word counts
for word, count in word_count.items():
        print(f"{word}\t{count}")
```