ASSIGNMENT 3

PROBLEM STATEMENT: 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.

CODE

wordcount_input.txt

```
hello hadoop
hadoop program
hadoop world
hello world
```

character_count.txt

```
abbc
ccba
aacb
```

mapper.py

```python
#!/usr/bin/env python3
import sys
for line in sys.stdin:
    for char in line.strip():
        print(f"{char}\t1")
```

reducer.py

```python
#!/usr/bin/env python3
import sys
from collections import defaultdict
counts = defaultdict(int)
for line in sys.stdin:
    key, val = line.strip().split("\t")
    counts[key] += int(val)

for key in sorted(counts):
    print(f"{key}\t{counts[key]}")
```

OUTPUT

```
hduser24@HDUSER-24865:~$ ls
hadoop-3.3.6.tar.gz   hadoop_lab_automation.sh  hadoop_tmp  input.txt        reducer.py
hadoop_automation.sh  hadoop_master.sh          hdfs        intermediate.txt
hduser24@HDUSER-24865:~$ mkdir hadoop_practicals
hduser24@HDUSER-24865:~$ cd hadoop_practicals/
hduser24@HDUSER-24865:~/hadoop_practicals$ mkdir input mapper reducer
hduser24@HDUSER-24865:~/hadoop_practicals$ cd input
hduser24@HDUSER-24865:~/hadoop_practicals/input$ nano wordcount_input.txt
hduser24@HDUSER-24865:~/hadoop_practicals/input$ start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [HDUSER-24865]
hduser24@HDUSER-24865:~/hadoop_practicals/input$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hduser24@HDUSER-24865:~/hadoop_practicals/input$ jps
8275 Jps
7894 NodeManager
7563 SecondaryNameNode
7197 NameNode
7325 DataNode
7774 ResourceManager
hduser24@HDUSER-24865:~/hadoop_practicals/input$ hdfs dfs -ls /
Found 3 items
drwxr-xr-x   - hduser24 supergroup          0 2025-04-19 07:36 /input
drwxr-xr-x   - hduser24 supergroup          0 2025-04-16 13:30 /tmp
drwxr-xr-x   - hduser24 supergroup          0 2025-04-16 13:30 /user
hduser24@HDUSER-24865:~/hadoop_practicals/input$ hdfs dfs -rm -r /input
Deleted /input
hduser24@HDUSER-24865:~/hadoop_practicals/input$ hdfs dfs -mkdir -p /input
```

```
hduser24@HDUSER-24865:~/hadoop_practicals/input$ hdfs dfs -ls /
Found 3 items
drwxr-xr-x   - hduser24 supergroup          0 2025-04-19 07:43 /input
drwxr-xr-x   - hduser24 supergroup          0 2025-04-16 13:30 /tmp
drwxr-xr-x   - hduser24 supergroup          0 2025-04-16 13:30 /user
hduser24@HDUSER-24865:~/hadoop_practicals/input$ hdfs dfs -put wordcount_input.txt /input/
hduser24@HDUSER-24865:~/hadoop_practicals/input$ hdfs dfs -ls /input/
Found 1 items
-rw-r--r--   1 hduser24 supergroup         53 2025-04-19 07:43 /input/wordcount_input.txt
hduser24@HDUSER-24865:~/hadoop_practicals/input$ cd ..
hduser24@HDUSER-24865:~/hadoop_practicals$ echo $HADOOP_HOME
/usr/local/hadoop/hadoop-3.3.6
hduser24@HDUSER-24865:~/hadoop_practicals$ hadoop jar
$HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.6.jar wordcount
/input /output
2025-04-19 07:44:30,524 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to
ResourceManager at /0.0.0.0:8032
2025-04-19 07:44:30,954 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path:
/tmp/hadoop-yarn/staging/hduser24/.staging/job_1745048577551_0001
2025-04-19 07:44:31,257 INFO input.FileInputFormat: Total input files to process : 1
2025-04-19 07:44:32,186 INFO mapreduce.JobSubmitter: number of splits:1
2025-04-19 07:44:32,817 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_1745048577551_0001
2025-04-19 07:44:32,818 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-04-19 07:44:33,068 INFO conf.Configuration: resource-types.xml not found
2025-04-19 07:44:33,070 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-04-19 07:44:33,369 INFO impl.YarnClientImpl: Submitted application
application_1745048577551_0001
2025-04-19 07:44:33,420 INFO mapreduce.Job: The url to track the job: http://HDUSER-
24865.:8088/proxy/application_1745048577551_0001/
2025-04-19 07:44:33,421 INFO mapreduce.Job: Running job: job_1745048577551_0001
2025-04-19 07:44:41,585 INFO mapreduce.Job: Job job_1745048577551_0001 running in uber mode
: false
2025-04-19 07:44:41,587 INFO mapreduce.Job:  map 0% reduce 0%
```

```
2025-04-19 07:44:46,711 INFO mapreduce.Job:  map 100% reduce 0%

2025-04-19 07:44:49,941 INFO mapreduce.Job:  map 100% reduce 100%

2025-04-19 07:44:51,985 INFO mapreduce.Job: Job job_1745048577551_0001 completed successfully

2025-04-19 07:44:52,118 INFO mapreduce.Job: Counters: 54

    File System Counters

        FILE: Number of bytes read=57

        FILE: Number of bytes written=554555

        FILE: Number of read operations=0

        FILE: Number of large read operations=0

        FILE: Number of write operations=0

        HDFS: Number of bytes read=165

        HDFS: Number of bytes written=35

        HDFS: Number of read operations=8

        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=1

        Launched reduce tasks=1

        Data-local map tasks=1

        Total time spent by all maps in occupied slots (ms)=2536

        Total time spent by all reduces in occupied slots (ms)=1703

        Total time spent by all map tasks (ms)=2536

        Total time spent by all reduce tasks (ms)=1703

        Total vcore-milliseconds taken by all map tasks=2536

        Total vcore-milliseconds taken by all reduce tasks=1703

        Total megabyte-milliseconds taken by all map tasks=1298432

        Total megabyte-milliseconds taken by all reduce tasks=871936

    Map-Reduce Framework

        Map input records=4
```

```
		Map output records=8
		Map output bytes=85
		Map output materialized bytes=57
		Input split bytes=112
		Combine input records=8
		Combine output records=4
		Reduce input groups=4
		Reduce shuffle bytes=57
		Reduce input records=4
		Reduce output records=4
		Spilled Records=8
		Shuffled Maps =1
		Failed Shuffles=0
		Merged Map outputs=1
		GC time elapsed (ms)=209
		CPU time spent (ms)=1560
		Physical memory (bytes) snapshot=732696576
		Virtual memory (bytes) snapshot=4237352960
		Total committed heap usage (bytes)=639631360
		Peak Map Physical memory (bytes)=474750976
		Peak Map Virtual memory (bytes)=2114686976
		Peak Reduce Physical memory (bytes)=257945600
		Peak Reduce Virtual memory (bytes)=2122665984
	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=53

        File Output Format Counters

                Bytes Written=35

hduser24@HDUSER-24865:~/hadoop_practicals$ hdfs dfs -ls /

Found 4 items

drwxr-xr-x   - hduser24 supergroup          0 2025-04-19 07:43 /input

drwxr-xr-x   - hduser24 supergroup          0 2025-04-19 07:44 /output

drwxr-xr-x   - hduser24 supergroup          0 2025-04-16 13:30 /tmp

drwxr-xr-x   - hduser24 supergroup          0 2025-04-16 13:30 /user

hduser24@HDUSER-24865:~/hadoop_practicals$ hdfs dfs -ls /output/

Found 2 items

-rw-r--r--   1 hduser24 supergroup          0 2025-04-19 07:44 /output/_SUCCESS

-rw-r--r--   1 hduser24 supergroup         35 2025-04-19 07:44 /output/part-r-00000

hduser24@HDUSER-24865:~/hadoop_practicals$ hdfs dfs -cat /output/part-r-00000

hadoop  3

hello   2

program 1

world   2

hduser24@HDUSER-24865:~/hadoop_practicals$ cd input

hduser24@HDUSER-24865:~/hadoop_practicals/input$ nano character_count.txt

hduser24@HDUSER-24865:~/hadoop_practicals/input$ hdfs dfs -put character_count.txt /input/

hduser24@HDUSER-24865:~/hadoop_practicals/input$ hdfs dfs -ls /input/

Found 2 items

-rw-r--r--   1 hduser24 supergroup         15 2025-04-19 07:50 /input/character_count.txt

-rw-r--r--   1 hduser24 supergroup         53 2025-04-19 07:43 /input/wordcount_input.txt

hduser24@HDUSER-24865:~/hadoop_practicals/input$ cd ..

hduser24@HDUSER-24865:~/hadoop_practicals$ nano ./reducer/reducer.py

hduser24@HDUSER-24865:~/hadoop_practicals$ nano ./mapper/mapper.py

hduser24@HDUSER-24865:~/hadoop_practicals$ chmod +x ./mapper/mapper.py

hduser24@HDUSER-24865:~/hadoop_practicals$ chmod +x reducer/reducer.py
```

```
hduser24@HDUSER-24865:~/hadoop_practicals$ hadoop jar
$HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.3.6.jar \

-input /input/character_count.txt \

-output /output/character_output \

-mapper mapper.py \

-reducer reducer.py \

-file ./mapper/mapper.py \

-file ./reducer/reducer.py

2025-04-19 08:21:01,496 WARN streaming.StreamJob: -file option is deprecated, please use generic
option -files instead.

packageJobJar: [./mapper/mapper.py, ./reducer/reducer.py, /tmp/hadoop-
unjar7467680124879837190/] [] /tmp/streamjob7901405455985332150.jar tmpDir=null

2025-04-19 08:21:02,401 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to
ResourceManager at /0.0.0.0:8032

2025-04-19 08:21:02,610 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to
ResourceManager at /0.0.0.0:8032

2025-04-19 08:21:03,055 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path:
/tmp/hadoop-yarn/staging/hduser24/.staging/job_1745048577551_0004

2025-04-19 08:21:03,390 INFO mapred.FileInputFormat: Total input files to process : 1

2025-04-19 08:21:03,894 INFO mapreduce.JobSubmitter: number of splits:3

2025-04-19 08:21:04,044 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_1745048577551_0004

2025-04-19 08:21:04,044 INFO mapreduce.JobSubmitter: Executing with tokens: []

2025-04-19 08:21:04,231 INFO conf.Configuration: resource-types.xml not found

2025-04-19 08:21:04,231 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.

2025-04-19 08:21:04,311 INFO impl.YarnClientImpl: Submitted application
application_1745048577551_0004

2025-04-19 08:21:04,361 INFO mapreduce.Job: The url to track the job: http://HDUSER-
24865.:8088/proxy/application_1745048577551_0004/

2025-04-19 08:21:04,364 INFO mapreduce.Job: Running job: job_1745048577551_0004

2025-04-19 08:21:10,479 INFO mapreduce.Job: Job job_1745048577551_0004 running in uber mode
: false

2025-04-19 08:21:10,482 INFO mapreduce.Job:  map 0% reduce 0%

2025-04-19 08:21:16,623 INFO mapreduce.Job:  map 100% reduce 0%
```

```
2025-04-19 08:21:21,681 INFO mapreduce.Job:  map 100% reduce 100%
2025-04-19 08:21:21,703 INFO mapreduce.Job: Job job_1745048577551_0004 completed successfully
2025-04-19 08:21:21,800 INFO mapreduce.Job: Counters: 54
    File System Counters
        FILE: Number of bytes read=78
        FILE: Number of bytes written=1123601
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=321
        HDFS: Number of bytes written=12
        HDFS: Number of read operations=14
        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=3
        Launched reduce tasks=1
        Data-local map tasks=3
        Total time spent by all maps in occupied slots (ms)=9375
        Total time spent by all reduces in occupied slots (ms)=2505
        Total time spent by all map tasks (ms)=9375
        Total time spent by all reduce tasks (ms)=2505
        Total vcore-milliseconds taken by all map tasks=9375
        Total vcore-milliseconds taken by all reduce tasks=2505
        Total megabyte-milliseconds taken by all map tasks=4800000
        Total megabyte-milliseconds taken by all reduce tasks=1282560
    Map-Reduce Framework
        Map input records=3
        Map output records=12
        Map output bytes=48
        Map output materialized bytes=90
        Input split bytes=297
        Combine input records=0
        Combine output records=0
        Reduce input groups=3
        Reduce shuffle bytes=90
        Reduce input records=12
        Reduce output records=3
        Spilled Records=24
        Shuffled Maps =3
        Failed Shuffles=0
        Merged Map outputs=3
        GC time elapsed (ms)=495
        CPU time spent (ms)=3270
        Physical memory (bytes) snapshot=1344417792
        Virtual memory (bytes) snapshot=8476753920
        Total committed heap usage (bytes)=1268252672
        Peak Map Physical memory (bytes)=473337856
        Peak Map Virtual memory (bytes)=2119614464
        Peak Reduce Physical memory (bytes)=242388992
        Peak Reduce Virtual memory (bytes)=2122412032
    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=24
    File Output Format Counters
        Bytes Written=12
2025-04-19 08:21:21,800 INFO streaming.StreamJob: Output directory: /output/character_output
hduser24@HDUSER-24865:~/hadoop_practicals$ hdfs dfs -ls /output/
Found 3 items
-rw-r--r--   1 hduser24 supergroup          0 2025-04-19 07:44 /output/_SUCCESS
drwxr-xr-x   - hduser24 supergroup          0 2025-04-19 08:21 /output/character_output
-rw-r--r--   1 hduser24 supergroup         35 2025-04-19 07:44 /output/part-r-00000
hduser24@HDUSER-24865:~/hadoop_practicals$ hdfs dfs -ls /output/character_output/
Found 2 items
-rw-r--r--   1 hduser24 supergroup          0 2025-04-19 08:21 /output/character_output/_SUCCESS
-rw-r--r--   1 hduser24 supergroup         12 2025-04-19 08:21 /output/character_output/part-00000
hduser24@HDUSER-24865:~/hadoop_practicals$ hdfs dfs -cat /output/character_output/part-00000
a       4
b       4
c       4
```