

EXP NO: 2

Run A Basic Word Count Map Reduce Program To Understand Map Reduce Paradigm

Input Files (wordCount.txt, mapper.py, reducer.py) :

```

Zellij (BA) [Tab #1]
soul@fedora:~/hadoop/input/pyscripts
soul@fedora:~/hadoop/input$ cd hadoop/input/
soul@fedora:~/hadoop/input$ ls
'DA Assignment 3' dataset.txt Experiments pig pyscripts shellscripts wordCount.txt
soul@fedora:~/hadoop/input$ cd pyscripts/
soul@fedora:~/hadoop/input/pyscripts$ ls
mapper-dataset.py mapper.py reducer-dataset.py reducer.py
soul@fedora:~/hadoop/input/pyscripts$ cat mapper.py
#!/usr/bin/python
import sys
for line in sys.stdin:
    line = line.strip()
    words = line.split()
    for word in words:
        print(f'{word}\t1')
soul@fedora:~/hadoop/input/pyscripts$

soul@fedora:~/hadoop/input/pyscripts
soul@fedora:~/hadoop/input$ cd hadoop/input/
soul@fedora:~/hadoop/input$ cd pyscripts/
soul@fedora:~/hadoop/input/pyscripts$ cd ..
soul@fedora:~/hadoop/input$ cat wordCount.txt
Oh I love it and I hate it at the same time
you and i drank the poison from the same wine
oh i love it and i hate it at the same time
hiding all of our sins from the daylight
from the daylight
running from the daylight
from the daylight
running from the daylight
soul@fedora:~/hadoop/input$ cd pyscripts/
soul@fedora:~/hadoop/input/pyscripts$ cat reducer.py
#!/usr/bin/python
from operator import itemgetter
import sys
current_word = None
word = None
current_count = 0
for line in sys.stdin:
    line = line.strip()
    word, count = line.split('\t', 1)
    count = int(count)
    if word == current_word:
        current_count += count
    else:
        if current_word:
            print(f'{current_word}\t{current_count}')
            current_count = count
            current_word = word
        if current_word == word:
            print(f'{current_word}\t{current_count}')
soul@fedora:~/hadoop/input/pyscripts$

```

```
$chmod 777 mapper.py reducer.py
```

```
$hadoop jar $HADOOP_STREAMING -input /exp2/word_count.txt -output
/exp2/output -mapper ~/DA-Lab/exp2/mapper.py -reducer ~/DA-
Lab/exp2/reducer.py
```

```

$ hadoop jar $HADOOP_STREAMING -input /exp1/word_count.txt -output /exp1/new_output -mapper ~/DA-Lab/mapper.py -reducer ~/DA-Lab/reducer.py
2024-08-28 03:19:55,058 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
packageJobJar: [/tmp/hadoop-unjar5640646781857695820/] [] /tmp/streamjob6319961755761707411.jar tmpDir=null
2024-08-28 03:19:58,598 INFO client.DefaultHadoopFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-08-28 03:19:59,548 INFO client.DefaultHadoopFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-08-28 03:20:02,347 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1724827394647_0002
2024-08-28 03:20:04,382 INFO mapred.FileInputFormat: Total input files to process : 1
2024-08-28 03:20:05,170 INFO mapreduce.JobSubmitter: number of splits:2
2024-08-28 03:20:06,083 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1724827394647_0002
2024-08-28 03:20:06,084 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-08-28 03:20:07,390 INFO conf.Configuration: resource-types.xml not found
2024-08-28 03:20:07,392 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-08-28 03:20:09,571 INFO impl.YarnClientImpl: Submitted application application_1724827394647_0002
2024-08-28 03:20:10,079 INFO mapreduce.Job: The url to track the job: http://kali:8088/proxy/application_1724827394647_0002/
2024-08-28 03:20:10,115 INFO mapreduce.Job: Running job: job_1724827394647_0002
2024-08-28 03:20:57,787 INFO mapreduce.Job: Job job_1724827394647_0002 running in uber mode : false
2024-08-28 03:20:57,794 INFO mapreduce.Job: map 0% reduce 0%
2024-08-28 03:21:27,334 INFO mapreduce.Job: map 100% reduce 0%
2024-08-28 03:21:50,144 INFO mapreduce.Job: map 100% reduce 100%

```

`$hdfs dfs -cat /exp2/output/*`

```
└─$ ./hdfs dfs -cat /exp2/output/*
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
2024-09-21 00:07:24,178 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
Hadoop 1
This 1
a 1
count 1
executed 1
in 1
is 1
program 1
word 1
```

Result :

Thus a basic Word Count program is run using map-reduce.