

Digesh Patel  
202318038  
Big Data Assignment-3

## Mapper.py

```
#!/usr/bin/python3 -0

import sys

# Loop through each line in the input
for line in sys.stdin:
    # Remove leading and trailing whitespace
    line = line.strip()
    # Split the line into words
    words = line.split()
    # Emit key-value pairs of word and count of 1
    for word in words:
        print(word, "\t", 1)

~
~
"mapper.py" 33L, 342B
```

**Shebang Line:**

- Specifies Python 3 interpreter.

**Importing sys Module:**

- sys module for system-specific functions.
- Provides access to system parameters.

**Loop through Input Lines:**

- Iterates over each line of input.
- Reads from standard input.

**Stripping Whitespace:**

- Removes leading and trailing whitespace.
- Ensures clean input.

**Splitting Lines into Words:**

- Splits lines into words.
- Based on whitespace.

**Emitting Key-Value Pairs:**

- Prints word and 1.
- Separated by a tab.

## Reducer.py

```
#!/usr/bin/python3 -O

import sys

# Initialize variables to keep track of current word and its count
current_word = None
current_count = 0

# Loop through each line in the input
for line in sys.stdin:
    # Split the line into word and count, separated by tab
    word, count = line.strip().split('\t', 1)

    # Convert count to integer
    count = int(count)

    # If the word is the same as the current word, increment its count
    if word == current_word:
        current_count += count
    else:
        # If the word is different, print the current word and its count
        if current_word:
            print(current_word, "\t", current_count)
        # Update current word and its count
        current_word = word
        current_count = count

# Print the last word and its count
if current_word:
    print(current_word, "\t", current_count)
~
~
"reducer.py" 33L, 863B
```

### Shebang Line:

- Specifies Python 3 interpreter.

### Importing sys Module:

- sys module for system-specific functions.
- Provides access to system parameters.

### Initialization:

- Initialize variables for word and count.
- current\_word and current\_count.

### Loop through Input Lines:

- Iterates over each line of input.
- Reads from standard input.

### Splitting Lines into Word and Count:

- Splits each line into word and count.
- Separated by tab, limiting to one split.

### Converting Count to Integer:

- Converts count from string to integer.
- Ensures numerical operations.

### Incrementing Word Count:

- Increments count if word is the same.
- Accumulates count for the same word.

### Printing Word Count:

- Prints current word and count.
- Separated by tab.

### Updating Current Word and Count:

- Updates current word and its count.
- Prepares for the next word.

### Printing Last Word and Count:

- Prints the last word and its count.
- Ensures all counts are accounted for.

## SingleNode , used txt file corpus.txt ~ 90 mb

The screenshot displays the AWS Management Console interface for a single EC2 instance. The instance is named '202318038-SingleNode' and is currently in a 'Running' state. The console shows various details including the instance ID, public IP address, private IP address, and the instance state. The instance is located in the 'ap-south-1' region.

Name	Instance ID	Instance state	Status check	Alarm status	Availabi...	Public I...	Public IPv4 ...	Elastic IP
202318038-SingleNode	i-08e9ceeb...	Running	2/2 checks pi	View alarms +	ap-south-1a	ec2-13-20...	13.201.125.239	-

**Instance: i-08e9ceebf5a9b46b (202318038-SingleNode)**

**Details** | Status and alarms New | Monitoring | Security | Networking | Storage | Tags

**Instance summary** Info

Instance ID	Public IPv4 address	Private IPv4 addresses
i-08e9ceebf5a9b46b (202318038-SingleNode)	13.201.125.239 <a href="#">open address</a>	172.31.41.195

IPv6 address: -

Instance state: Running

Public IPv4 DNS: ec2-13-201-125-239.ap-south-1.compute.amazonaws.com [open address](#)

## Without Hadoop

```
ubuntu@ip-172-31-41-195:~$ time cat corpus.txt | python3 mapper.py | sort | python3 reducer.py
!          26
!!!!!!    1
!)         1
"          482
" "        14
" ", "pc"   7
" ", "pcs"  7
" "Do       1
" "The      1
" "There's  1
```

```
real    1m6.636s
user    0m5.413s
sys     0m0.466s
ubuntu@ip-172-31-41-195:~$ |
```

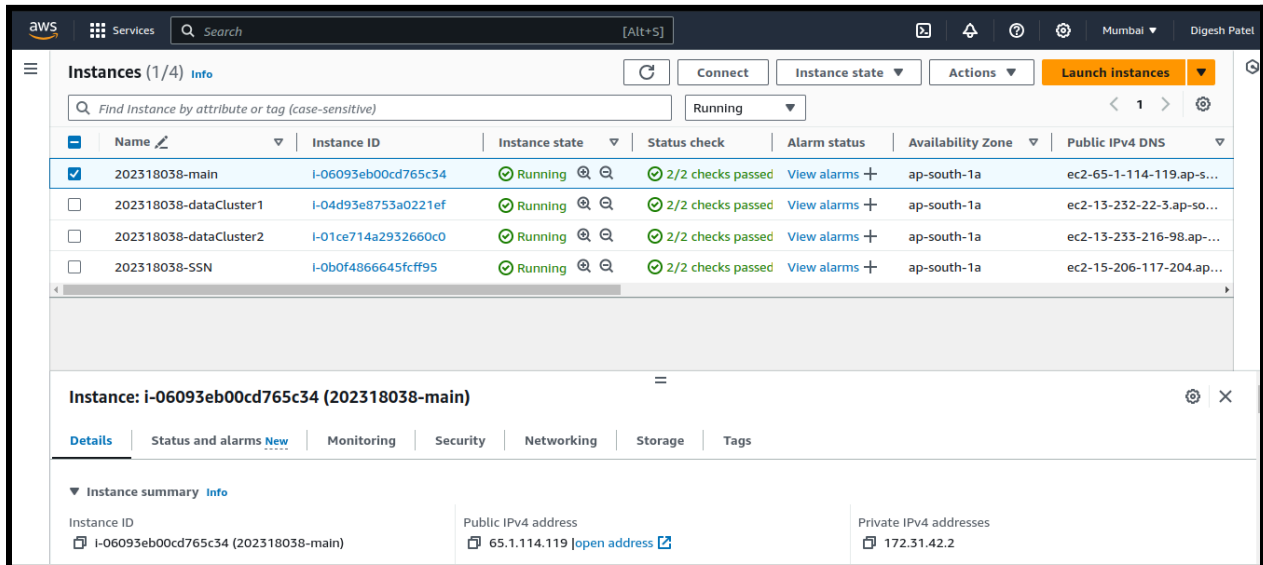
## With Hadoop

```
ubuntu@ip-172-31-41-195:~$ time hadoop jar /home/ubuntu/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.9.1.jar -mapper /home/ubuntu/mapper.py -reducer /home/ubuntu/reducer.py -input /input/corpus.txt -output /output/wordcounts
packageJobJar: [/tmp/hadoop-unjar5082568089063517402/] [] /tmp/streamjob8271078526609092725.jar tmpDir=null
24/02/24 05:08:12 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
24/02/24 05:08:12 INFO mapred.FileInputFormat: Total input files to process : 1
24/02/24 05:08:13 INFO mapreduce.JobSubmitter: number of splits:2
24/02/24 05:08:13 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
24/02/24 05:08:13 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1708748829675_0005
24/02/24 05:08:14 INFO impl.YarnClientImpl: Submitted application application_1708748829675_0005
24/02/24 05:08:14 INFO mapreduce.Job: The url to track the job: http://ip-172-31-41-195.ap-south-1.compute.internal:8088/proxy/application_1708748829675_0005/
24/02/24 05:08:14 INFO mapreduce.Job: Running job: job_1708748829675_0005
24/02/24 05:08:20 INFO mapreduce.Job: Job job_1708748829675_0005 running in uber mode : false
24/02/24 05:08:20 INFO mapreduce.Job:  map 0% reduce 0%
24/02/24 05:08:38 INFO mapreduce.Job:  map 37% reduce 0%
24/02/24 05:08:44 INFO mapreduce.Job:  map 56% reduce 0%
24/02/24 05:08:50 INFO mapreduce.Job:  map 67% reduce 0%
24/02/24 05:08:55 INFO mapreduce.Job:  map 83% reduce 0%
24/02/24 05:08:56 INFO mapreduce.Job:  map 100% reduce 0%
24/02/24 05:09:12 INFO mapreduce.Job:  map 100% reduce 94%
24/02/24 05:09:15 INFO mapreduce.Job:  map 100% reduce 100%
24/02/24 05:09:16 INFO mapreduce.Job: Job job_1708748829675_0005 completed successfully
24/02/24 05:09:16 INFO mapreduce.Job: Counters: 49
    File System Counters
      FILE: Number of bytes read=341906164
      FILE: Number of bytes written=513458624
      FILE: Number of read operations=0
      FILE: Number of large read operations=0
      FILE: Number of write operations=0
      HDFS: Number of bytes read=87466386
      HDFS: Number of bytes written=7297521
      HDFS: Number of read operations=9
      HDFS: Number of large read operations=0
      HDFS: Number of write operations=2
```

```
real    2m17.928s
user    2m13.154s
sys     0m3.162s
ubuntu@ip-172-31-41-195:~$ |
```

```
ubuntu@ip-172-31-41-195:~$ hdfs dfs -cat /output/wordcounts/part-00000
!                26
!!!!!!!         1
!)              1
"              482
" "            14
" " ,"pc"       7
" " ,"pcs"      7
""Do           1
""The          1
""There's      1
""We           2
""},           7
"#[FROM        1
"#[FROM        1
```

## MultiNode , used txt file corpus.txt ~ 90 mb



**Instances (1/4)** Info

Find Instance by attribute or tag (case-sensitive)

Running

Name	Instance ID	Instance state	Status check	Alarm status	Availability Zone	Public IPv4 DNS
202318038-main	i-06093eb00cd765c34	Running	2/2 checks passed	View alarms	ap-south-1a	ec2-65-1-114-119.ap-s...
202318038-dataCluster1	i-04d93e8753a0221ef	Running	2/2 checks passed	View alarms	ap-south-1a	ec2-13-232-22-3.ap-so...
202318038-dataCluster2	i-01ce714a2932660c0	Running	2/2 checks passed	View alarms	ap-south-1a	ec2-13-233-216-98.ap...
202318038-SSN	i-0b0f4866645fcff95	Running	2/2 checks passed	View alarms	ap-south-1a	ec2-15-206-117-204.ap...

**Instance: i-06093eb00cd765c34 (202318038-main)**

Details | Status and alarms New | Monitoring | Security | Networking | Storage | Tags

▼ Instance summary Info

Instance ID i-06093eb00cd765c34 (202318038-main)	Public IPv4 address 65.1.114.119 <a href="#">open address</a>	Private IPv4 addresses 172.31.42.2
---	--	---------------------------------------

## With Hadoop

```
ubuntu@ip-172-31-42-2:~/hadoop$ sbin/start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [ip-172-31-42-2.ap-south-1.compute.internal]
ip-172-31-42-2.ap-south-1.compute.internal: starting namenode, logging to /home/ubuntu/hadoop/logs/hadoop-ubuntu-namenod
172.31.42.2: starting datanode, logging to /home/ubuntu/hadoop/logs/hadoop-ubuntu-datanode-ip-172-31-42-2.out
172.31.44.214: starting datanode, logging to /home/ubuntu/hadoop/logs/hadoop-ubuntu-datanode-ip-172-31-44-214.out
172.31.32.110: starting datanode, logging to /home/ubuntu/hadoop/logs/hadoop-ubuntu-datanode-ip-172-31-32-110.out
```

```
ubuntu@ip-172-31-42-2:~$ hadoop jar /home/ubuntu/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.9.1.jar -mapper /home/ubuntu/mapper.py -reducer /home/ubuntu/reducer.py -input /input/corpus.txt -output /output/wordcounts
packageJobJar: [/tmp/hadoop-unjar1936485731669817299/] [] /tmp/streamjob7987625646435085579.jar tmpDir=null
24/02/24 10:53:14 INFO client.RMProxy: Connecting to ResourceManager at /172.31.42.2:8032
24/02/24 10:53:14 INFO client.RMProxy: Connecting to ResourceManager at /172.31.42.2:8032
24/02/24 10:53:15 INFO mapred.FileInputFormat: Total input files to process : 1
24/02/24 10:53:15 INFO mapreduce.JobSubmitter: number of splits:2
24/02/24 10:53:15 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
24/02/24 10:53:15 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1708769543844_0002
24/02/24 10:53:15 INFO impl.YarnClientImpl: Submitted application application_1708769543844_0002
24/02/24 10:53:15 INFO mapreduce.Job: The url to track the job: http://ip-172-31-42-2.ap-south-1.compute.internal:8088/proxy/application_1708769543844_0002/
24/02/24 10:53:15 INFO mapreduce.Job: Running job: job_1708769543844_0002
24/02/24 10:53:22 INFO mapreduce.Job: Job job_1708769543844_0002 running in uber mode : false
24/02/24 10:53:22 INFO mapreduce.Job: map 0% reduce 0%
24/02/24 10:53:40 INFO mapreduce.Job: map 37% reduce 0%
24/02/24 10:53:46 INFO mapreduce.Job: map 58% reduce 0%
24/02/24 10:53:52 INFO mapreduce.Job: map 67% reduce 0%
24/02/24 10:53:57 INFO mapreduce.Job: map 100% reduce 0%
24/02/24 10:54:14 INFO mapreduce.Job: map 100% reduce 94%
24/02/24 10:54:16 INFO mapreduce.Job: map 100% reduce 100%
24/02/24 10:54:18 INFO mapreduce.Job: Job job_1708769543844_0002 completed successfully
```

```
real    0m37.713s
user    0m32.432s
sys     0m4.448s
ubuntu@ip-172-31-42-2:~$
```

```
ubuntu@ip-172-31-42-2:~$ hdfs dfs -ls /output/wordcounts
Found 2 items
-rw-r--r--  3 ubuntu supergroup          0 2024-02-24 10:54 /output/wordcounts/_SUCCESS
-rw-r--r--  3 ubuntu supergroup 7297521 2024-02-24 10:54 /output/wordcounts/part-00000
ubuntu@ip-172-31-42-2:~$ hdfs dfs -cat /output/wordcounts/part-00000
!          26
!!!!!!!    1
!)         1
"         482
"         14
"         7
"         7
"         1
"         1
"         1
"         1
"         1
```