BIG DATA ANALYTICS

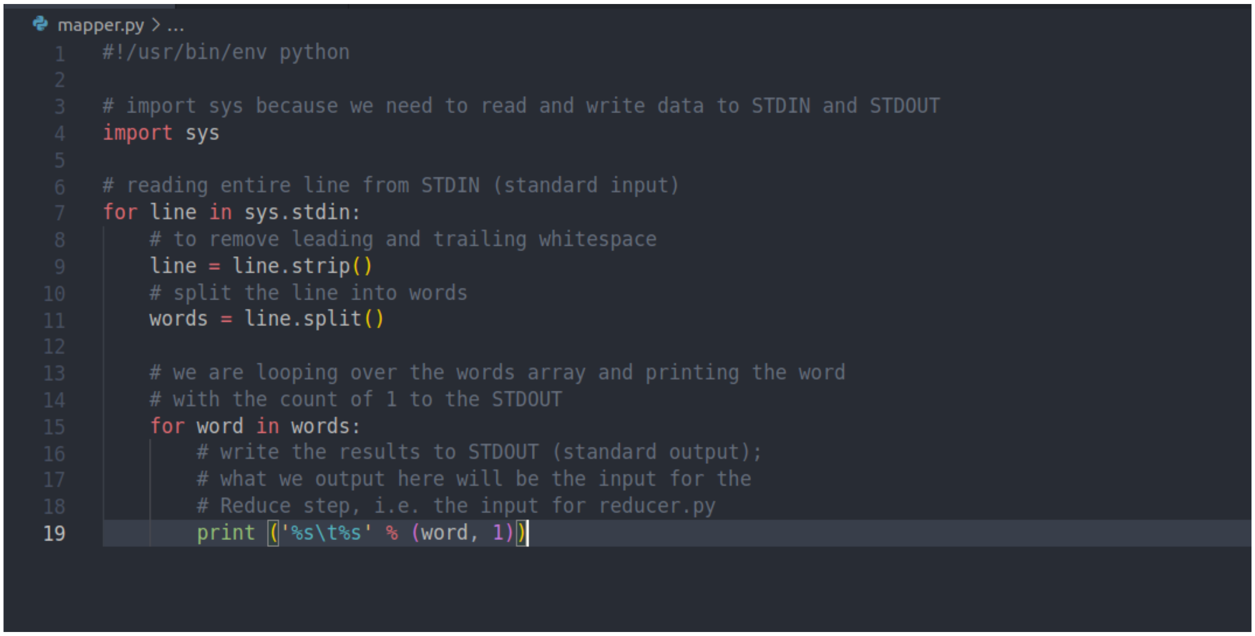
Practical 6

Analyse impact of different number of mapper and reducer on same definition as practical 4.

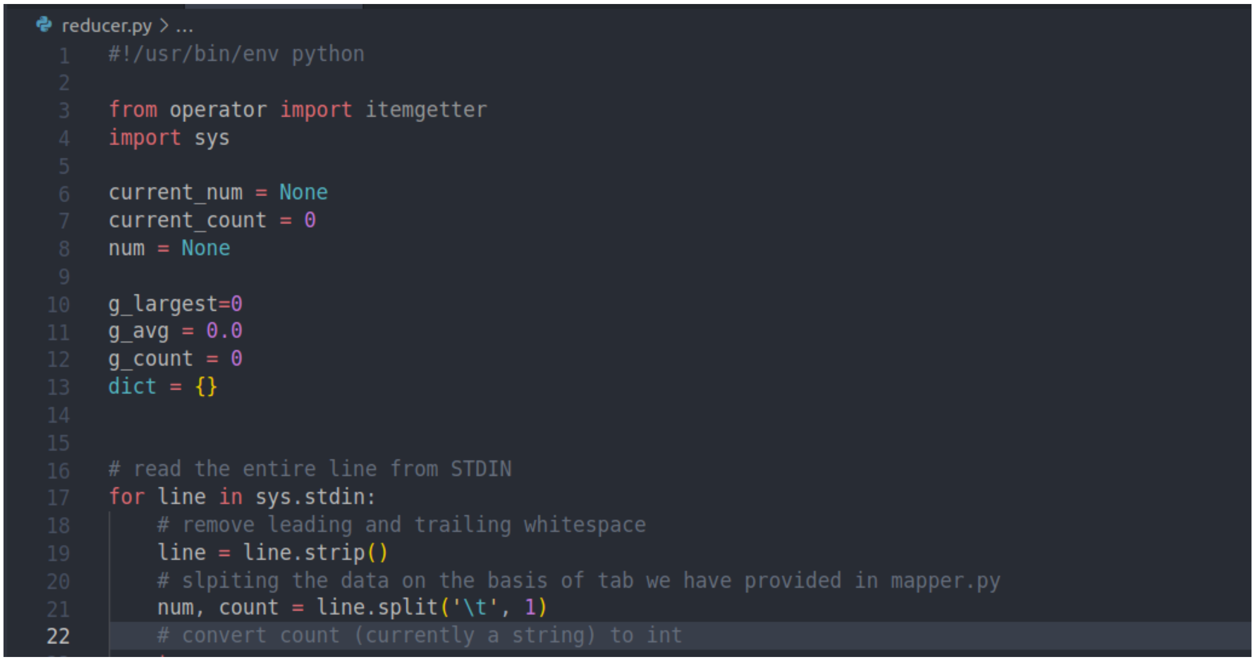
Prepared By

18BCE151 - Parth Kapadia

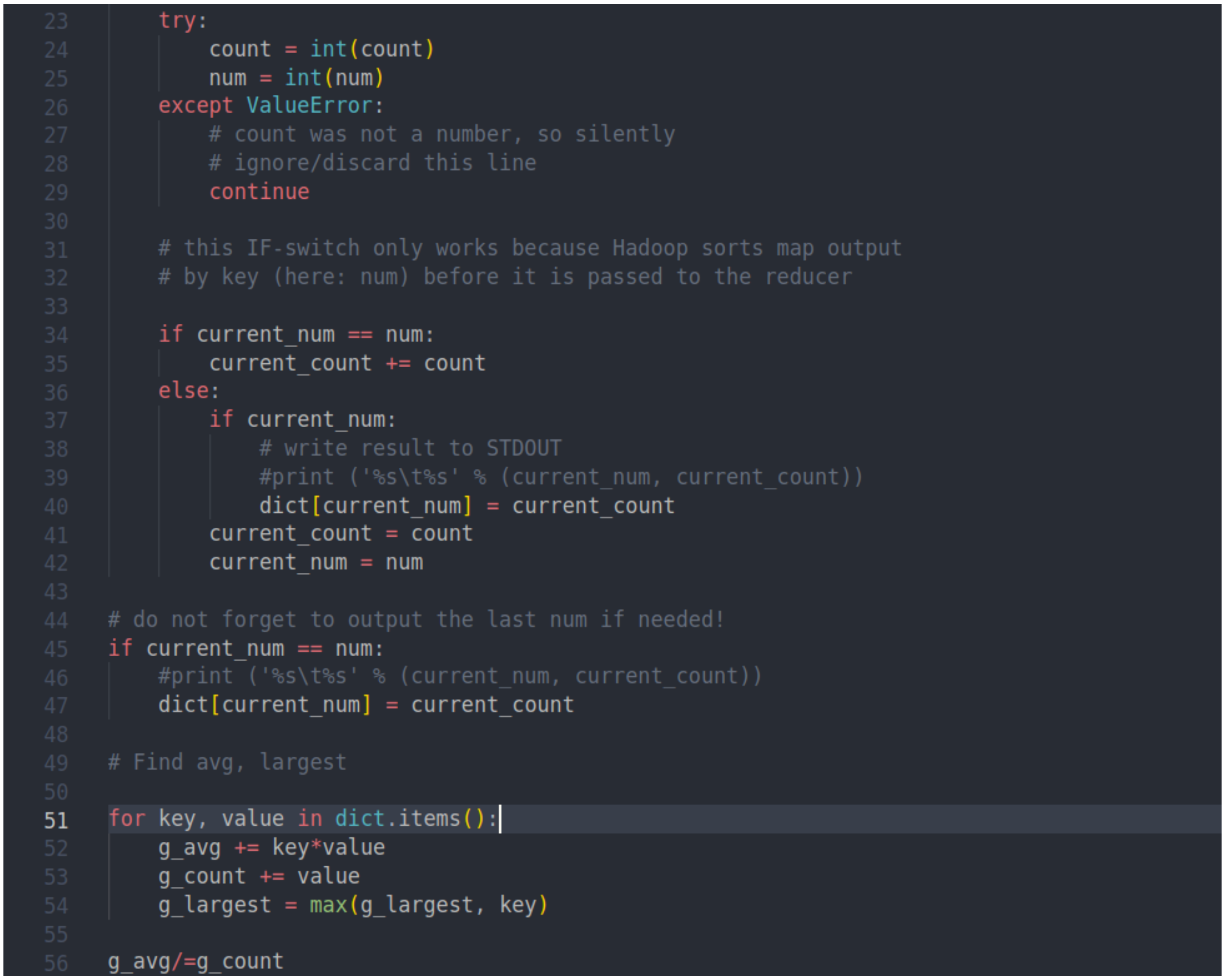
Creating Mapper and Reducer



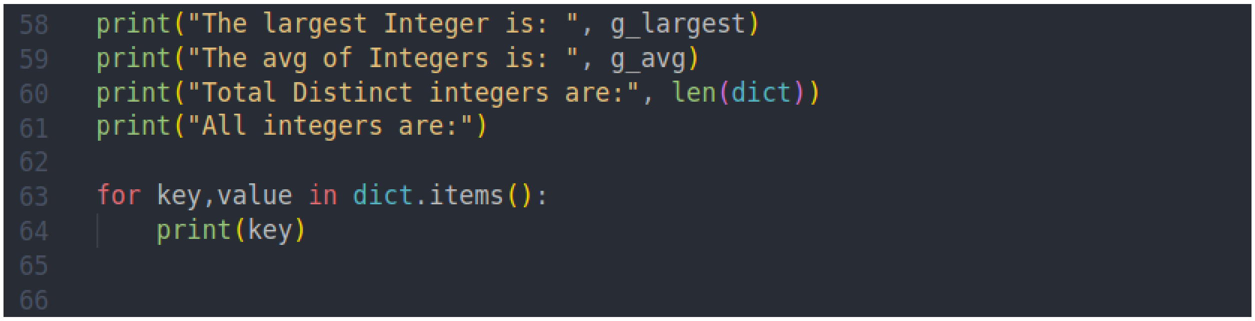
```python
#!/usr/bin/env python

# import sys because we need to read and write data to STDIN and STDOUT
import sys

# reading entire line from STDIN (standard input)
for line in sys.stdin:
    # to remove leading and trailing whitespace
    line = line.strip()
    # split the line into words
    words = line.split()

    # we are looping over the words array and printing the word
    # with the count of 1 to the STDOUT
    for word in words:
        # write the results to STDOUT (standard output);
        # what we output here will be the input for the
        # Reduce step, i.e. the input for reducer.py
        print ('%s\t%s' % (word, 1))
```



```python
#!/usr/bin/env python

from operator import itemgetter
import sys

current_num = None
current_count = 0
num = None

g_largest=0
g_avg = 0.0
g_count = 0
dict = {}


# read the entire line from STDIN
for line in sys.stdin:
    # remove leading and trailing whitespace
    line = line.strip()
    # slpiting the data on the basis of tab we have provided in mapper.py
    num, count = line.split('\t', 1)
    # convert count (currently a string) to int
```



```
    try:
        count = int(count)
        num = int(num)
    except ValueError:
        # count was not a number, so silently
        # ignore/discard this line
        continue

    # this IF-switch only works because Hadoop sorts map output
    # by key (here: num) before it is passed to the reducer

    if current_num == num:
        current_count += count
    else:
        if current_num:
            # write result to STDOUT
            #print ('%s\t%s' % (current_num, current_count))
            dict[current_num] = current_count
        current_count = count
        current_num = num

# do not forget to output the last num if needed!
if current_num == num:
    #print ('%s\t%s' % (current_num, current_count))
    dict[current_num] = current_count

# Find avg, largest

for key, value in dict.items():
    g_avg += key*value
    g_count += value
    g_largest = max(g_largest, key)

g_avg/=g_count
```



```
print("The largest Integer is: ", g_largest)
print("The avg of Integers is: ", g_avg)
print("Total Distinct integers are:", len(dict))
print("All integers are:")

for key,value in dict.items():
    print(key)
```

Running the Map Reduce program

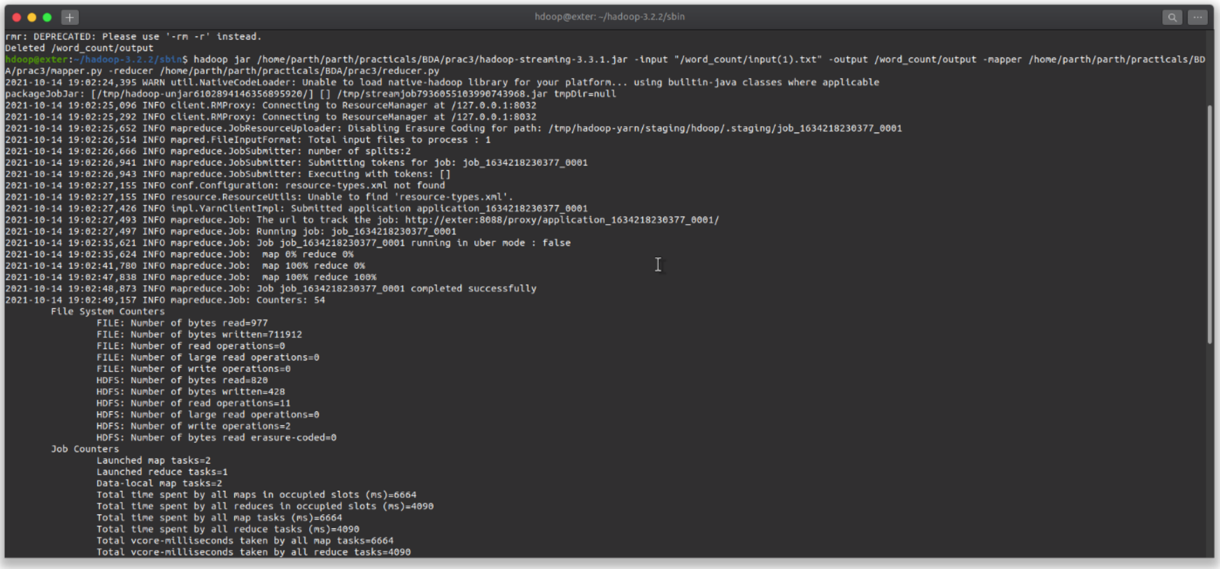
hadoop jar /home/parth/parth/practicals/BDA/prac3/hadoop-streaming-3.3.1.jar

-input "/word_count/input(1).txt"

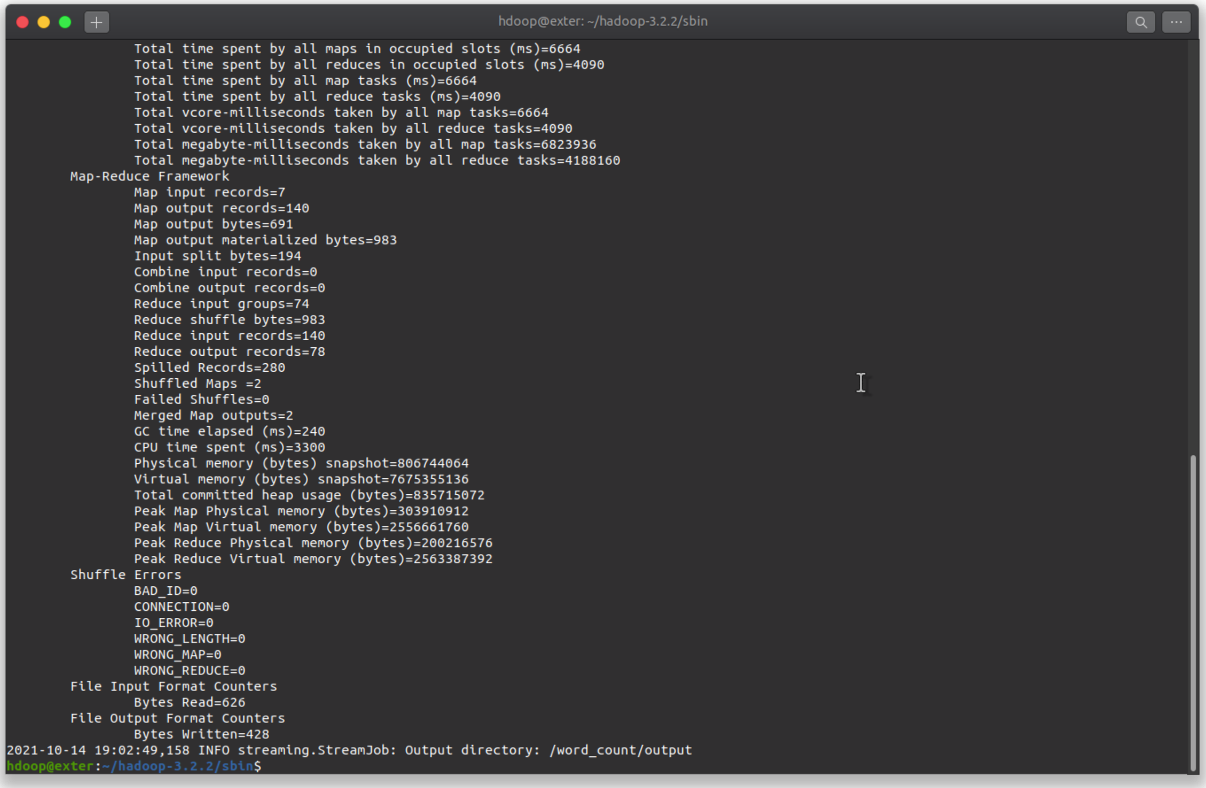
-output /word_count/output

-mapper /home/parth/parth/practicals/BDA/prac3/mapper.py

-reducer /home/parth/parth/practicals/BDA/prac3/reducer.py



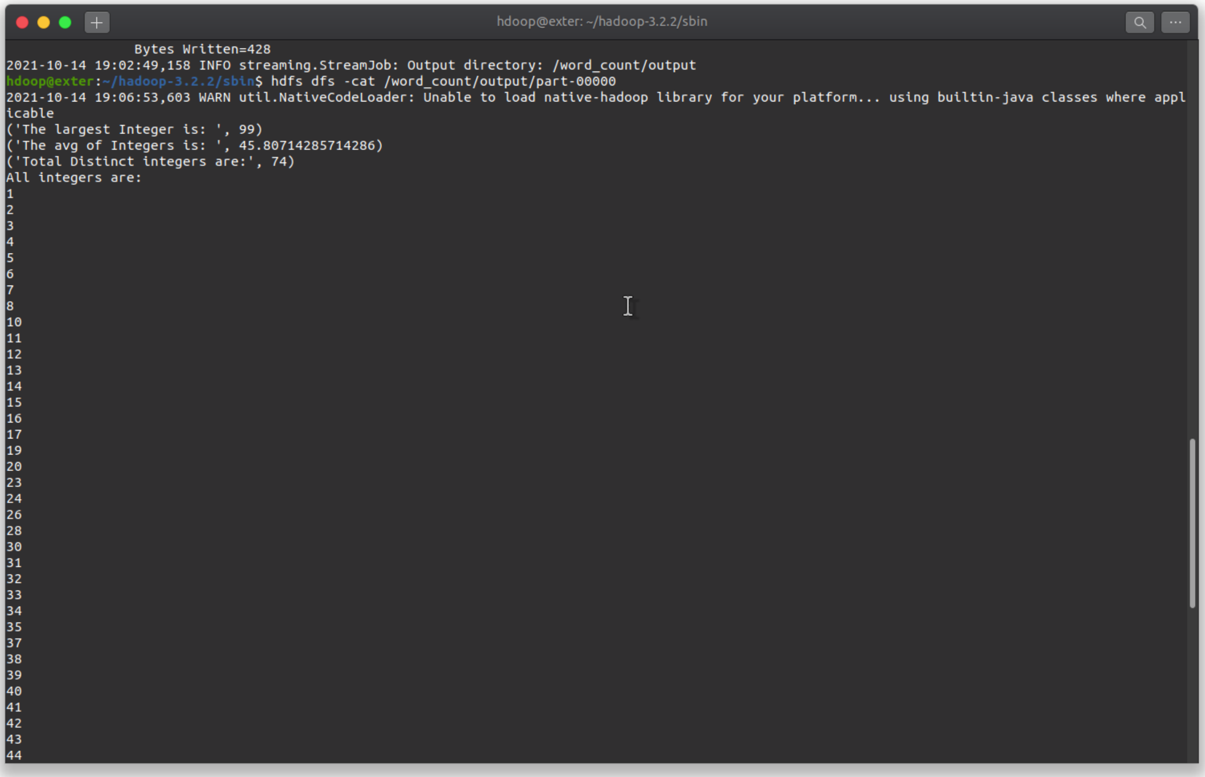
```
rmr: DEPRECATED: Please use '-rm -r' instead.
Deleted /word_count/output
hdoop@exter:~/hadoop-3.2.2/sbin$ hadoop jar /home/parth/parth/practicals/BDA/prac3/hadoop-streaming-3.3.1.jar -input "/word_count/input(1).txt" -output /word_count/output -mapper /home/parth/parth/practicals/BD
A/prac3/mapper.py -reducer /home/parth/parth/practicals/BDA/prac3/reducer.py
2021-10-14 19:02:24,395 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
packageJobJar: [/tmp/hadoop-unjar6102894146356895920/] [] /tmp/streamjob7936055103990743968.jar tmpDir=null
2021-10-14 19:02:25,096 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
2021-10-14 19:02:25,292 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
2021-10-14 19:02:25,652 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hdoop/.staging/job_1634218230377_0001
2021-10-14 19:02:26,514 INFO mapred.FileInputFormat: Total input files to process : 1
2021-10-14 19:02:26,666 INFO mapreduce.JobSubmitter: number of splits:2
2021-10-14 19:02:26,941 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1634218230377_0001
2021-10-14 19:02:26,943 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-10-14 19:02:27,155 INFO conf.Configuration: resource-types.xml not found
2021-10-14 19:02:27,155 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-10-14 19:02:27,426 INFO impl.YarnClientImpl: Submitted application application_1634218230377_0001
2021-10-14 19:02:27,493 INFO mapreduce.Job: The url to track the job: http://exter:8088/proxy/application_1634218230377_0001/
2021-10-14 19:02:27,497 INFO mapreduce.Job: Running job: job_1634218230377_0001
2021-10-14 19:02:35,621 INFO mapreduce.Job: Job job_1634218230377_0001 running in uber mode : false
2021-10-14 19:02:35,624 INFO mapreduce.Job:  map 0% reduce 0%
2021-10-14 19:02:41,780 INFO mapreduce.Job:  map 100% reduce 0%
2021-10-14 19:02:47,838 INFO mapreduce.Job:  map 100% reduce 100%
2021-10-14 19:02:48,873 INFO mapreduce.Job: Job job_1634218230377_0001 completed successfully
2021-10-14 19:02:49,157 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=977
                FILE: Number of bytes written=711912
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=820
                HDFS: Number of bytes written=428
                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)=6664
                Total time spent by all reduces in occupied slots (ms)=4090
                Total time spent by all map tasks (ms)=6664
                Total time spent by all reduce tasks (ms)=4090
                Total vcore-milliseconds taken by all map tasks=6664
                Total vcore-milliseconds taken by all reduce tasks=4090
```



```
                Total time spent by all maps in occupied slots (ms)=6664
                Total time spent by all reduces in occupied slots (ms)=4090
                Total time spent by all map tasks (ms)=6664
                Total time spent by all reduce tasks (ms)=4090
                Total vcore-milliseconds taken by all map tasks=6664
                Total vcore-milliseconds taken by all reduce tasks=4090
                Total megabyte-milliseconds taken by all map tasks=6823936
                Total megabyte-milliseconds taken by all reduce tasks=4188160
        Map-Reduce Framework
                Map input records=7
                Map output records=140
                Map output bytes=691
                Map output materialized bytes=983
                Input split bytes=194
                Combine input records=0
                Combine output records=0
                Reduce input groups=74
                Reduce shuffle bytes=983
                Reduce input records=140
                Reduce output records=78
                Spilled Records=280
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=240
                CPU time spent (ms)=3300
                Physical memory (bytes) snapshot=806744064
                Virtual memory (bytes) snapshot=7675355136
                Total committed heap usage (bytes)=835715072
                Peak Map Physical memory (bytes)=303910912
                Peak Map Virtual memory (bytes)=2556661760
                Peak Reduce Physical memory (bytes)=200216576
                Peak Reduce Virtual memory (bytes)=2563387392
        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=626
        File Output Format Counters
                Bytes Written=428
2021-10-14 19:02:49,158 INFO streaming.StreamJob: Output directory: /word_count/output
hdoop@exter:~/hadoop-3.2.2/sbin$
```

Output

hdfs dfs -cat /word_count/output/part-00000



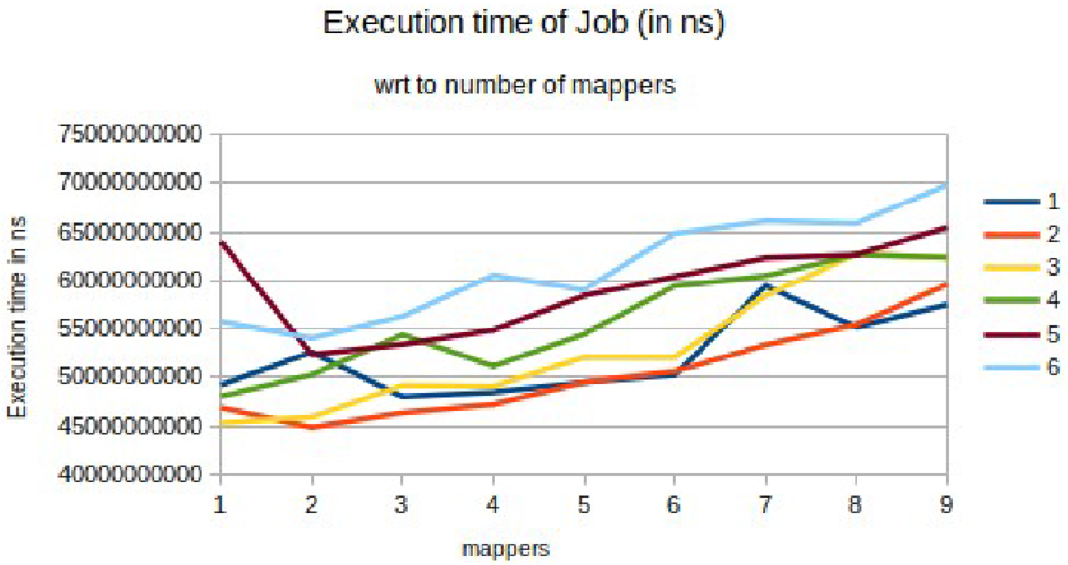
```
                Bytes Written=428
2021-10-14 19:02:49,158 INFO streaming.StreamJob: Output directory: /word_count/output
hdoop@exter:~/hadoop-3.2.2/sbin$ hdfs dfs -cat /word_count/output/part-00000
2021-10-14 19:06:53,603 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where appl
icable
('The largest Integer is: ', 99)
('The avg of Integers is: ', 45.80714285714286)
('Total Distinct integers are:', 74)
All integers are:
1
2
3
4
5
6
7
8
10
11
12
13
14
15
16
17
19
20
23
24
26
28
30
31
32
33
34
35
37
38
39
40
41
42
43
44
```

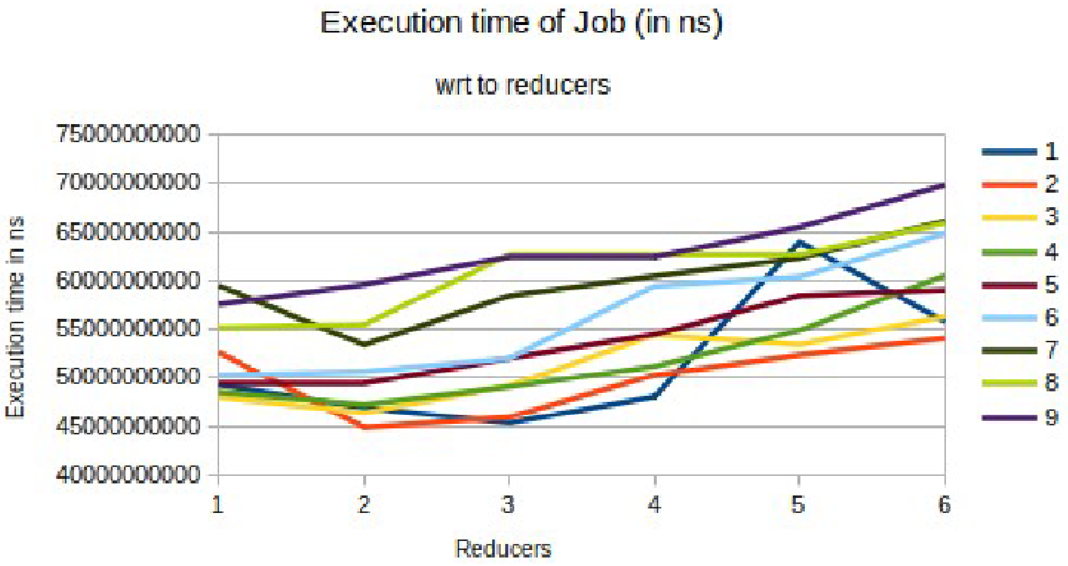
Changing number of mappers and reducers

-Dmapreduce.job.reduces=3 -Dmapreduce.job.maps=2

hadoop jar /home/parth/parth/practicals/BDA/prac3/hadoop-streaming-3.3.1.jar -Dmapreduce.job.reduces=3 -Dmapreduce.job.maps=2 -input /word_count/Essay700kB.txt -output /word_count/output -mapper

```
/home/parth/parth/practicals/BDA/prac6/mapper.py -reducer
/home/parth/parth/practicals/BDA/prac6/reducer.py
```


Execution time of Job (in ns)
wrt to number of mappers
Execution time in ns
mappers


Execution time of Job (in ns)
wrt to reducers
Execution time in ns
Reducers