Module 2: Assignment - 4

1. Find out the count of each word in the ‘Shakespeare.txt’ dataset in the ‘Shakespeare.rar’

Solution:

Reading Shakespeare.txt

```
bitnami@debian:~$ hadoop fs -cat /user/bitnami/Shakespeare.txt_
```

```
  Par. Faith sir, ha's led the drumme before the English
Tragedians: to belye him I will not, and more of his
souldiership I know not, except in that Country, he had
the honour to be the Officer at a place there called Mile-end,
to instruct for the doubling of files. I would doe the
man what honour I can, but of this I am not certaine

   Cap.G. He hath out-villain'd villanie so farre, that the
raritie redeemes him

   Ber. A pox on him, he's a Cat still

   Int. His qualities being at this poore price, I neede
not to aske you, if Gold will corrupt him to reuolt

   Par. Sir, for a Cardceue he will sell the fee-simple of
his saluation, the inheritance of it, and cut th' intaile from
all remainders, and a perpetuall succession for it perpetually

   Int. What's his Brother, the other Captain Dumain?
  Cap.E. Why do's he aske him of me?
  Int. What's he?
  Par. E'ne a Crow a'th same nest: not altogether so
great as the first in goodnesse, but greater a great deale in
```

Running Map Reduce Job:

```
bitnami@debian:~$ hadoop jar map_reduce_3.jar map_reduce_3/WordCount /user/bitnami/Shakespeare.txt
 /user/bitnami/output_wordcount_1
2021-11-04 18:29:24,654 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceMana
ger at /0.0.0.0:8032
Exception in thread "main" org.apache.hadoop.mapred.FileAlreadyExistsException: Output directory h
dfs://localhost:8020/user/bitnami/output_wordcount_1 already exists
        at org.apache.hadoop.mapreduce.lib.output.FileOutputFormat.checkOutputSpecs(FileOutputForm
at.java:164)
        at org.apache.hadoop.mapreduce.JobSubmitter.checkSpecs(JobSubmitter.java:277)
        at org.apache.hadoop.mapreduce.JobSubmitter.submitJobInternal(JobSubmitter.java:143)
        at org.apache.hadoop.mapreduce.Job$11.run(Job.java:1571)
        at org.apache.hadoop.mapreduce.Job$11.run(Job.java:1568)
        at java.security.AccessController.doPrivileged(Native Method)
        at javax.security.auth.Subject.doAs(Subject.java:422)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1878)
        at org.apache.hadoop.mapreduce.Job.submit(Job.java:1568)
        at org.apache.hadoop.mapreduce.Job.waitForCompletion(Job.java:1589)
        at map_reduce_3.WordCount.main(WordCount.java:94)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:323)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:236)
bitnami@debian:~$ hadoop jar map_reduce_3.jar map_reduce_3/WordCount /user/bitnami/Shakespeare.txt
 /user/bitnami/out_wordcount_1
2021-11-04 18:29:59,595 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceMana
ger at /0.0.0.0:8032
2021-11-04 18:30:00,034 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not
 performed. Implement the Tool interface and execute your application with ToolRunner to remedy th
is.
2021-11-04 18:30:00,052 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tm
p/hadoop-yarn/staging/hadoop/.staging/job_1636027398107_0002
2021-11-04 18:30:00,327 INFO input.FileInputFormat: Total input files to process : 1
2021-11-04 18:30:00,812 INFO mapreduce.JobSubmitter: number of splits:1
2021-11-04 18:30:01,053 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1636027398107_
0002
2021-11-04 18:30:01,053 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-11-04 18:30:01,276 INFO conf.Configuration: resource-types.xml not found
2021-11-04 18:30:01,276 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-11-04 18:30:01,352 INFO impl.YarnClientImpl: Submitted application application_1636027398107_
0002
2021-11-04 18:30:01,415 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/
application_1636027398107_0002/
2021-11-04 18:30:01,416 INFO mapreduce.Job: Running job: job_1636027398107_0002
2021-11-04 18:30:08,588 INFO mapreduce.Job: Job job_1636027398107_0002 running in uber mode : fals
e
2021-11-04 18:30:08,591 INFO mapreduce.Job:  map 0% reduce 0%
2021-11-04 18:30:15,735 INFO mapreduce.Job:  map 100% reduce 0%
2021-11-04 18:30:22,797 INFO mapreduce.Job:  map 100% reduce 100%
2021-11-04 18:30:23,836 INFO mapreduce.Job: Job job_1636027398107_0002 completed successfully
```

Viewing count for each word using two reducers only output:

```
bitnami@debian:~$ hadoop fs -ls /user/bitnami/out_wordcount_1/
Found 2 items
-rw-r--r--   1 hadoop supergroup          0 2021-11-04 18:30 /user/bitnami/out_wordcount_1/_SUCCESS
-rw-r--r--   1 hadoop supergroup     746266 2021-11-04 18:30 /user/bitnami/out_wordcount_1/part-r-00000
```

```
bitnami@debian:~$ hadoop fs -cat /user/bitnami/out_wordcount_1/part-r-00000
```

```
willingly       24
willingly,      4
willingly.      1
willingly:      3
willingnesse    1
willingnesse.   1
willow  2
wills   6
wills,  4
wills.  1
wills;  1
wils    1
wils:   1
wils?   1
wilt    215
wilt,   11
wilt.   4
wilt:   2
wilt;   1
wilte   1
wimpled,        1
win     56
win,    4
win.    2
win;    1
winch,  1
winch:  1
winck   1
wincke, 1
wind    7
wind,   3
wind-pipes      1
wind-swift      1
wind:   2
winde   74
winde)  1
winde,  25
winde-obeying   1
winde-shak'd-Surge,     1
winde-shaken.   1
```

```
yrefull 1
yron    5
yron:   2
ysickle 1
ysicles,        1
yssue   4
yssue,  1
yssue:  1
yssued, 1
yssues. 1
yt      2
yut     1
z       1
zeale   11
zeale,  7
zeale:  1
zeale?  1
zeales, 1
zealous 3
zelous  2
zip     1
zo      1
zwaggerd        1
```

3. Find out the most commonly used words (Words with the count over 100 are considered common).

Solution:

Running Map Reduce Job

```
bitnami@debian:~$ hadoop jar map_reduce_4.jar map_reduce_3/WordCount /user/bitnami/Shakespeare.txt /user/bitnami/output_wordcount_5
```

```
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Killed reduce tasks=1
                Launched map tasks=1
                Launched reduce tasks=2
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=9224
                Total time spent by all reduces in occupied slots (ms)=16240
                Total time spent by all map tasks (ms)=4612
                Total time spent by all reduce tasks (ms)=8120
                Total vcore-milliseconds taken by all map tasks=4612
                Total vcore-milliseconds taken by all reduce tasks=8120
                Total megabyte-milliseconds taken by all map tasks=9445376
                Total megabyte-milliseconds taken by all reduce tasks=16629760
        Map-Reduce Framework
                Map input records=135281
                Map output records=817811
                Map output bytes=7680316
                Map output materialized bytes=1020341
                Input split bytes=115
                Combine input records=817811
                Combine output records=70503
                Reduce input groups=70503
                Reduce shuffle bytes=1020341
                Reduce input records=70503
                Reduce output records=806
                Spilled Records=141006
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=214
                CPU time spent (ms)=6980
                Physical memory (bytes) snapshot=807919616
                Virtual memory (bytes) snapshot=10361499648
                Total committed heap usage (bytes)=645398528
                Peak Map Physical memory (bytes)=366923776
                Peak Map Virtual memory (bytes)=3451076608
                Peak Reduce Physical memory (bytes)=221237248
                Peak Reduce Virtual memory (bytes)=3456024576
        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=4516586
        File Output Format Counters
                Bytes Written=7701
```

Viewing output for most commonly used words (count > 100)

```
bitnami@debian:~$ hadoop fs -cat /user/bitnami/output_wordcount_5/part-r-00000_
```

```
'Tis    523
'tis    692
1       202
A       1513
Against 126
Ah      102
All     327
All.    108
An      127
An.     139
And     7029
Ant.    450
Art     105
At      203
Be      411
Before  106
Being   101
Ber.    217
Bru.    200
Brutus  101
But     2326
By      679
Caesar  191
Caesar, 107
Cassi.  108
Clau.   100
Cleo.   210
Court   103
Did     218
Do      277
Du.     108
Duke    444
Duke.   205
England 111
Euen    216
Exeunt. 639
Fal.    327
Fathers 188
For     1636
Fortune 103
France  126
France, 154
Gentleman       122
Gods    238
Goe     132
Good    425
Had     137
Hath    274
Haue    410
```

```
euen    280
euer    538
eye     143
face    173
faire   464
fall    196
false   177
feare   350
feare,  102
fellow  111
fit     123
fiue    100
follow  204
for     5372
forth   231
foule   162
foure   108
friends 113
full    334
gentle  287
goe     432
gone    173
gone,   129
good    1984
good,   106
gracious        158
great   688
ha's    160
had     1194
hand    350
hand,   248
hard    118
hath    1541
haue    5023
haue,   104
he      4051
he's    103
he,     108
heare   665
hearts  117
heere   623
heere,  163
hence   131
hence,  123
her     2759
her,    373
her:    127
him?    135
himselfe        233
```

```
then,   244
these   1000
things  224
thinke  748
this?   187
those   411
though  356
thought 271
through 201
thy     3488
till    383
time    613
time,   222
to      14978
told    198
truth   124
verie   118
vnder   178
vnto    339
vpon    1240
vs      1086
vs,     266
warrant 137
was     1856
way     330
way,    112
wee     218
were    1263
where   615
which   1220
while   135
whom    285
wife    139
wife,   100
wish    169
within  193
words   203
words,  105
worthy  148
would   1882
wrong   122
y       133
yong    197
you     9838
you,    1043
you.    213
you:    281
your    6186
youth   103
```