# Big Data Workshop Week-6

**Pratik Sarkar 2039301**

1. **Word Count Version 1**

a. Compile the file

```
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v1$ javac -classpath $
(hadoop classpath) WordCount.java
```

b. Produce the jar file

```
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v1$ jar cf wordcount.j
ar Word*.class
```

c. Create the input directory

```
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v1$ hdfs dfs -mkdir in
put_word
2021-03-19 23:47:36,190 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
```

d. Create the input files

```
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v1$ echo A long time a
go in a galaxy far far away > testfile1
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v1$ echo Another episo
de of Star Wars > testfile2
```

e. Saves the files to input directory

```
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v1$ hdfs dfs -put test
file? input_word
2021-03-19 23:48:41,016 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
```

f. Delete the previous output directory

```
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v1$ hdfs dfs -rm -R ou
tput_word
2021-03-19 23:48:55,439 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
Deleted output_word
```

g. Run the map reduce program

```
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v1$ hadoop jar wordcou
nt.jar WordCount input_word output_word
2021-03-19 23:54:36,145 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
2021-03-19 23:54:36,876 INFO client.RMProxy: Connecting to ResourceManager at /
127.0.0.1:8032
2021-03-19 23:54:37,296 WARN mapreduce.JobResourceUploader: Hadoop command-line
 option parsing not performed. Implement the Tool interface and execute your ap
plication with ToolRunner to remedy this.
2021-03-19 23:54:37,316 INFO mapreduce.JobResourceUploader: Disabling Erasure C
oding for path: /tmp/hadoop-yarn/staging/pratik/.staging/job_1616166771362_0009
2021-03-19 23:54:37,580 INFO input.FileInputFormat: Total input files to proces
s : 2
2021-03-19 23:54:37,654 INFO mapreduce.JobSubmitter: number of splits:2
2021-03-19 23:54:37,810 INFO mapreduce.JobSubmitter: Submitting tokens for job:
 job_1616166771362_0009
2021-03-19 23:54:37,812 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-03-19 23:54:38,026 INFO conf.Configuration: resource-types.xml not found
2021-03-19 23:54:38,027 INFO resource.ResourceUtils: Unable to find 'resource-t
ypes.xml'.
2021-03-19 23:54:38,103 INFO impl.YarnClientImpl: Submitted application applica
tion_1616166771362_0009
2021-03-19 23:54:38,147 INFO mapreduce.Job: The url to track the job: http://pr
atik-VirtualBox:8088/proxy/application_1616166771362_0009/
2021-03-19 23:54:38,147 INFO mapreduce.Job: Running job: job_1616166771362_0009
2021-03-19 23:54:45,259 INFO mapreduce.Job: Job job_1616166771362_0009 running
in uber mode : false
2021-03-19 23:54:45,260 INFO mapreduce.Job:  map 0% reduce 0%
2021-03-19 23:54:50.393 INFO mapreduce.Job:  map 100% reduce 0%
                Reduce shuffle bytes=162
                Reduce input records=14
                Reduce output records=14
                Spilled Records=28
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=238
                CPU time spent (ms)=1860
                Physical memory (bytes) snapshot=732311552
                Virtual memory (bytes) snapshot=7769075712
                Total committed heap usage (bytes)=597164032
                Peak Map Physical memory (bytes)=298344448
                Peak Map Virtual memory (bytes)=2586935296
                Peak Reduce Physical memory (bytes)=195788800
                Peak Reduce Virtual memory (bytes)=2595938304
        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=70
        File Output Format Counters
                Bytes Written=94
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v1$
```

h. Checking files in output directory

```
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v1$ hdfs dfs -ls outpu
t_word
2021-03-19 23:55:58,989 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r--   1 pratik supergroup          0 2021-03-19 23:54 output_word/_SUCCE
SS
-rw-r--r--   1 pratik supergroup         94 2021-03-19 23:54 output_word/part-r
-00000
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v1$
```

i. See what is in the output file

```
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v1$ hdfs dfs -cat outp
ut_word/part-r-00000
2021-03-19 23:56:26,861 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
A       1
Another 1
Star    1
Wars    1
a       1
ago     1
away    1
episode 1
far     2
galaxy  1
in      1
long    1
of      1
time    1
```

j. Retrieving the result

```
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v1$ hdfs dfs -get outp
ut_word/part-r-00000 word-results.txt
2021-03-19 23:57:14,471 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v1$
```

### 1.1 Using the larger dataset

a. Removing previous test files

```
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v1$ hdfs dfs -rm input
_word/testfile?
2021-03-19 23:58:08,992 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
Deleted input_word/testfile1
Deleted input_word/testfile2
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v1$
```

b. Save new file to input directory

```
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v1$ hdfs dfs -put shak
espeare.txt input_word
2021-03-19 23:58:33,860 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v1$
```

c. Delete previous output directory

```
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v1$ hdfs dfs -rm -R ou
tput_word
2021-03-19 23:59:03,135 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
Deleted output_word
```

d. Run the Map Reduce program

```
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v1$ hadoop jar wordcou
nt.jar WordCount input_word output_word
2021-03-19 23:59:29,993 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
2021-03-19 23:59:30,744 INFO client.RMProxy: Connecting to ResourceManager at /
127.0.0.1:8032
2021-03-19 23:59:31,183 WARN mapreduce.JobResourceUploader: Hadoop command-line
 option parsing not performed. Implement the Tool interface and execute your ap
plication with ToolRunner to remedy this.
2021-03-19 23:59:31,232 INFO mapreduce.JobResourceUploader: Disabling Erasure C
oding for path: /tmp/hadoop-yarn/staging/pratik/.staging/job_1616166771362_0010
2021-03-19 23:59:31,479 INFO input.FileInputFormat: Total input files to proces
s : 1
2021-03-19 23:59:31,558 INFO mapreduce.JobSubmitter: number of splits:1
2021-03-19 23:59:31,715 INFO mapreduce.JobSubmitter: Submitting tokens for job:
 job_1616166771362_0010
2021-03-19 23:59:31,717 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-03-19 23:59:31,974 INFO conf.Configuration: resource-types.xml not found
2021-03-19 23:59:31,974 INFO resource.ResourceUtils: Unable to find 'resource-t
ypes.xml'.
2021-03-19 23:59:32,050 INFO impl.YarnClientImpl: Submitted application applica
tion_1616166771362_0010
2021-03-19 23:59:32,093 INFO mapreduce.Job: The url to track the job: http://pr
atik-VirtualBox:8088/proxy/application_1616166771362_0010/
2021-03-19 23:59:32,094 INFO mapreduce.Job: Running job: job_1616166771362_0010
2021-03-19 23:59:39,701 INFO mapreduce.Job: Job job_1616166771362_0010 running
in uber mode : false
2021-03-19 23:59:39,703 INFO mapreduce.Job:  map 0% reduce 0%
2021-03-19 23:59:48.860 INFO mapreduce.Job:  map 100% reduce 0%
                Reduce shuffle bytes=978916
                Reduce input records=67505
                Reduce output records=67505
                Spilled Records=135010
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=243
                CPU time spent (ms)=6020
                Physical memory (bytes) snapshot=697921536
                Virtual memory (bytes) snapshot=5184892928
                Total committed heap usage (bytes)=676855808
                Peak Map Physical memory (bytes)=497651712
                Peak Map Virtual memory (bytes)=2587963392
                Peak Reduce Physical memory (bytes)=200269824
                Peak Reduce Virtual memory (bytes)=2596929536
        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=5458200
        File Output Format Counters
                Bytes Written=717768
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v1$
```

e. Check files in the output directory

```
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v1$ hdfs dfs -ls outpu
t_word
2021-03-20 00:01:20,204 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r--   1 pratik supergroup          0 2021-03-19 23:59 output_word/_SUCCE
SS
-rw-r--r--   1 pratik supergroup     717768 2021-03-19 23:59 output_word/part-r
-00000
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v1$
```

f. See what is in the output file

```
something,      4
something-      1
something-settled       1
something.      10
something;      7
sometime        56
sometime's      1
sometime,       2
sometimes       40
sometimes,      6
sometimes:      1
sometimes?      1
somever 1
somewhat        11
somewhat.       2
somewhere       3
son     318
son!    25
son!'?  1
son'    2
son'-   1
son';   1
son's   16
son's,  1
son's.  1
son,    150
son-    11
son-in-law      4
son-in-law's    1
```

g. Copy the file locally

```
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v1$ hdfs dfs -get outp
ut_word/part-r-00000 shakespeare-results.txt
2021-03-20 00:02:31,858 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v1$
```

h. Using more to view results

```
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v1$ more shakespeare-r
esults.txt
"       241
"'Tis   1
"A      4
"AS-IS".        1
"Air,"  1
"Alas,  1
"Amen"  2
"Amen"? 1
"Amen," 1
"And    1
"Aroint 1
"B      1
"Black  1
"Break  1
"Brutus"        1
"Brutus,        2
"C      1
"Caesar"?       1
"Caesar,        1
"Caesar."       2
"Certes,"       1
"Come   1
"Cursed 1
"D      1
"Darest 1
"Defect"        1
```

## 2. Word Count Version 2

a. Using Grep to search

```
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v1$ grep anon shakespe
are-results.txt
Canonized,      1
abroad-anon     1
anon    30
anon!   2
anon,   12
anon-   2
anon.   41
anon.-  1
anon;   5
anon?   1
anonymous       1
canon   2
canon,  1
canon.  2
canon;  1
canoniz'd       1
canoniz'd,      1
canonize        1
canonized       1
canons, 1
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v1$ grep Anon shakespe
are-results.txt
'Anon   1
'Anon!' 1
'Anon,  1
Anon    9
```

**Making Wordcount Case Insensitive:**

b. Compile the file and create a Jar file

```
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v2$ javac -classpath $
(hadoop classpath) WordCount.java
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v2$ jar cf wordcount.j
ar Word*.class
```

c. Removing output directory

```
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v2$ hdfs dfs -rm -R ou
tput_word
2021-03-19 21:31:34,054 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
Deleted output_word
```

d. Running the program

```
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v2$ hadoop jar wordcou
nt.jar WordCount input_word output_word
2021-03-19 21:31:46,598 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
2021-03-19 21:31:47,245 INFO client.RMProxy: Connecting to ResourceManager at /
127.0.0.1:8032
2021-03-19 21:31:47,626 WARN mapreduce.JobResourceUploader: Hadoop command-line
 option parsing not performed. Implement the Tool interface and execute your ap
plication with ToolRunner to remedy this.
2021-03-19 21:31:47,648 INFO mapreduce.JobResourceUploader: Disabling Erasure C
oding for path: /tmp/hadoop-yarn/staging/pratik/.staging/job_1616166771362_0004
2021-03-19 21:31:47,906 INFO input.FileInputFormat: Total input files to proces
s : 1
2021-03-19 21:31:48,414 INFO mapreduce.JobSubmitter: number of splits:1
2021-03-19 21:31:48,559 INFO mapreduce.JobSubmitter: Submitting tokens for job:
 job_1616166771362_0004
2021-03-19 21:31:48,560 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-03-19 21:31:48,763 INFO conf.Configuration: resource-types.xml not found
2021-03-19 21:31:48,763 INFO resource.ResourceUtils: Unable to find 'resource-t
ypes.xml'.
2021-03-19 21:31:48,829 INFO impl.YarnClientImpl: Submitted application applica
tion_1616166771362_0004
2021-03-19 21:31:48,874 INFO mapreduce.Job: The url to track the job: http://pr
atik-VirtualBox:8088/proxy/application_1616166771362_0004/
2021-03-19 21:31:48,875 INFO mapreduce.Job: Running job: job_1616166771362_0004
2021-03-19 21:31:55,056 INFO mapreduce.Job: Job job_1616166771362_0004 running
in uber mode : false
2021-03-19 21:31:55,058 INFO mapreduce.Job:  map 0% reduce 0%
2021-03-19 21:32:02.183 INFO mapreduce.Job:  map 100% reduce 0%
                Reduce input groups=59508
                Reduce shuffle bytes=869870
                Reduce input records=59508
                Reduce output records=59508
                Spilled Records=119016
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=215
                CPU time spent (ms)=5530
                Physical memory (bytes) snapshot=699281408
                Virtual memory (bytes) snapshot=5181775872
                Total committed heap usage (bytes)=705691648
                Peak Map Physical memory (bytes)=499273728
                Peak Map Virtual memory (bytes)=2586107904
                Peak Reduce Physical memory (bytes)=200007680
                Peak Reduce Virtual memory (bytes)=2595667968
        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=5458200
        File Output Format Counters
                Bytes Written=640196
```

```
2021-03-19 21:32:07,236 INFO mapreduce.Job:  map 100% reduce 100%
2021-03-19 21:32:07,259 INFO mapreduce.Job: Job job_1616166771362_0004 complete
d successfully
2021-03-19 21:32:07,360 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=869870
                FILE: Number of bytes written=2208249
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=5458325
                HDFS: Number of bytes written=640196
                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)=4812
                Total time spent by all reduces in occupied slots (ms)=2503
                Total time spent by all map tasks (ms)=4812
                Total time spent by all reduce tasks (ms)=2503
                Total vcore-milliseconds taken by all map tasks=4812
                Total vcore-milliseconds taken by all reduce tasks=2503
                Total megabyte-milliseconds taken by all map tasks=4927488
                Total megabyte-milliseconds taken by all reduce tasks=2563072
        Map-Reduce Framework
```

e. Retrieving the results file from hdfs

```
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v2$ hdfs dfs -get  out
put_word/part-r-00000 shakespeare-results.txt
2021-03-19 21:32:13,687 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
```

f. Using Grep to search again

```
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v2$ grep anon shakespe
are-results.txt
'anon   1
'anon!' 1
'anon,  1
abroad-anon     1
anon    39
anon!   3
anon,   30
anon-   2
anon.   43
anon.-  1
anon;   5
anon?   1
anonymous       1
canon   2
canon,  1
canon.  2
canon;  1
canoniz'd       1
canoniz'd,      1
canonize        1
canonized       1
canonized,      1
canons, 1
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v2$ grep Anon shakespe
are-results.txt
```

### 3. Word Count Version 3

Removing punctuations from word:

a. Compile the file and create a Jar file

```
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v3$ javac -classpath $
(hadoop classpath) WordCount.java
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v3$ jar cf wordcount.j
ar Word*.class
```

b. Removing output directory

```
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v3$ hdfs dfs -rm -R ou
tput_word
2021-03-19 21:44:55,552 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
Deleted output_word
```

c. Running the program

```
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v3$ hadoop jar wordcou
nt.jar WordCount input_word output_word
2021-03-19 21:45:12,588 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
2021-03-19 21:45:13,378 INFO client.RMProxy: Connecting to ResourceManager at /
127.0.0.1:8032
2021-03-19 21:45:13,747 WARN mapreduce.JobResourceUploader: Hadoop command-line
 option parsing not performed. Implement the Tool interface and execute your ap
plication with ToolRunner to remedy this.
2021-03-19 21:45:13,779 INFO mapreduce.JobResourceUploader: Disabling Erasure C
oding for path: /tmp/hadoop-yarn/staging/pratik/.staging/job_1616166771362_0006
2021-03-19 21:45:13,991 INFO input.FileInputFormat: Total input files to proces
s : 1
2021-03-19 21:45:14,531 INFO mapreduce.JobSubmitter: number of splits:1
2021-03-19 21:45:14,667 INFO mapreduce.JobSubmitter: Submitting tokens for job:
 job_1616166771362_0006
2021-03-19 21:45:14,668 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-03-19 21:45:14,865 INFO conf.Configuration: resource-types.xml not found
2021-03-19 21:45:14,866 INFO resource.ResourceUtils: Unable to find 'resource-t
ypes.xml'.
2021-03-19 21:45:14,935 INFO impl.YarnClientImpl: Submitted application applica
tion_1616166771362_0006
2021-03-19 21:45:14,977 INFO mapreduce.Job: The url to track the job: http://pr
atik-VirtualBox:8088/proxy/application_1616166771362_0006/
2021-03-19 21:45:14,977 INFO mapreduce.Job: Running job: job_1616166771362_0006
2021-03-19 21:45:21,105 INFO mapreduce.Job: Job job_1616166771362_0006 running
in uber mode : false
2021-03-19 21:45:21,106 INFO mapreduce.Job:  map 0% reduce 0%
2021-03-19 21:45:29,256 INFO mapreduce.Job:  map 100% reduce 0%
```

```
2021-03-19 21:45:34,310 INFO mapreduce.Job:  map 100% reduce 100%
2021-03-19 21:45:34,327 INFO mapreduce.Job: Job job_1616166771362_0006 complete
d successfully
2021-03-19 21:45:34,424 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=334066
                FILE: Number of bytes written=1136991
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=5458325
                HDFS: Number of bytes written=245910
                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)=5792
                Total time spent by all reduces in occupied slots (ms)=2295
                Total time spent by all map tasks (ms)=5792
                Total time spent by all reduce tasks (ms)=2295
                Total vcore-milliseconds taken by all map tasks=5792
                Total vcore-milliseconds taken by all reduce tasks=2295
                Total megabyte-milliseconds taken by all map tasks=5931008
                Reduce shuffle bytes=334066
                Reduce input records=23722
                Reduce output records=23722
                Spilled Records=47444
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=239
                CPU time spent (ms)=6800
                Physical memory (bytes) snapshot=785809408
                Virtual memory (bytes) snapshot=5178654720
                Total committed heap usage (bytes)=667942912
                Peak Map Physical memory (bytes)=593219584
                Peak Map Virtual memory (bytes)=2586931200
                Peak Reduce Physical memory (bytes)=192589824
                Peak Reduce Virtual memory (bytes)=2591723520
        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=5458200
        File Output Format Counters
                Bytes Written=245910
```

d. Retrieving the results file from hdfs

```
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v3$ hdfs dfs -get  out
put_word/part-r-00000 shakespeare-results.csv
2021-03-19 21:50:02,234 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
```

e. Using Grep to search

```
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v3$ grep anon shakespe
are-results.csv
anon,128
anonymous,1
canon,6
canoniz,2
canonize,1
canonized,2
canons,1
```

The results are still case insensitive

```
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/wordcount-v3$ grep Anon shakespe
are-results.csv
```

## 4. Updated Word Count

Amended Wordcount file to remove some more punctuation marks:

```
        if (word.contains("!") || word.contains("[") || word.contains("\"") ||
word.contains("]"))
            continue;
        if (word.contains("?") || word.contains("*") || word.contains(".") ||
word.contains("#"))
            continue;
        if (word.contains("_") || word.contains("@") || word.contains(",") ||
word.contains("&"))
            continue;
        //Updated Word Count to check for additional punctuations.
        if (word.contains(":") || word.contains(";") || word.contains("'") ||
word.contains("-"))
            continue;
```

## 5. Python

a. Giving permission to mapper and reducer

```
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2$ chmod +x $HADOOP_HOME/week6/pyt
hon/mapper.py
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2$ chmod +x $HADOOP_HOME/week6/pyt
hon/reducer.py
```

b. Test the mapper

```
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2$ echo "foo foo quux labs foo bar
 quux" | $HADOOP_HOME/week6/python/mapper.py
foo     1
foo     1
quux    1
labs    1
foo     1
bar     1
quux    1
```

c. Test the mapper and reducer together

```
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2$ echo "foo foo quux labs foo bar
 quux" | $HADOOP_HOME/week6/python/mapper.py | sort -k1,1 | $HADOOP_HOME/week6/
python/reducer.py
bar     1
foo     3
labs    1
quux    2
```

d. Running the mapper and reducer together in HDFS

```
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2$ hadoop jar /home/pratik/hadoop/
hadoop-3.2.2/share/hadoop/tools/lib/hadoop-streaming-3.2.2.jar -mapper $HADOOP_
HOME/week6/python/mapper.py -reducer $HADOOP_HOME/week6/python/reducer.py -inpu
t input_word/shakespeare.txt -output output_word
2021-03-20 00:26:30,428 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
packageJobJar: [/tmp/hadoop-unjar2942593522301913971/] [] /tmp/streamjob5144539
032253817474.jar tmpDir=null
2021-03-20 00:26:31,109 INFO client.RMProxy: Connecting to ResourceManager at /
127.0.0.1:8032
2021-03-20 00:26:31,313 INFO client.RMProxy: Connecting to ResourceManager at /
127.0.0.1:8032
2021-03-20 00:26:31,507 INFO mapreduce.JobResourceUploader: Disabling Erasure C
oding for path: /tmp/hadoop-yarn/staging/pratik/.staging/job_1616166771362_0012
2021-03-20 00:26:31,735 INFO mapred.FileInputFormat: Total input files to proce
ss : 1
2021-03-20 00:26:31,808 INFO mapreduce.JobSubmitter: number of splits:2
2021-03-20 00:26:31,927 INFO mapreduce.JobSubmitter: Submitting tokens for job:
 job_1616166771362_0012
2021-03-20 00:26:31,929 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-03-20 00:26:32,110 INFO conf.Configuration: resource-types.xml not found
2021-03-20 00:26:32,110 INFO resource.ResourceUtils: Unable to find 'resource-t
ypes.xml'.
2021-03-20 00:26:32,174 INFO impl.YarnClientImpl: Submitted application applica
tion_1616166771362_0012
2021-03-20 00:26:32,218 INFO mapreduce.Job: The url to track the job: http://pr
atik-VirtualBox:8088/proxy/application_1616166771362_0012/
2021-03-20 00:26:32,219 INFO mapreduce.Job: Running job: job_1616166771362_0012
2021-03-20 00:26:38,871 INFO mapreduce.Job:  map 0% reduce 0%
2021-03-20 00:26:45,987 INFO mapreduce.Job:  map 50% reduce 0%
2021-03-20 00:26:47,003 INFO mapreduce.Job:  map 100% reduce 0%
2021-03-20 00:26:54,083 INFO mapreduce.Job:  map 100% reduce 100%
2021-03-20 00:26:54,112 INFO mapreduce.Job: Job job_1616166771362_0012 complete
d successfully
2021-03-20 00:26:54,195 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=8546440
                FILE: Number of bytes written=17803009
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=5462520
                HDFS: Number of bytes written=717768
                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)=12121
                Total time spent by all reduces in occupied slots (ms)=4866
                Total time spent by all map tasks (ms)=12121
                Total time spent by all reduce tasks (ms)=4866
                Total vcore-milliseconds taken by all map tasks=12121
```

```
                Reduce input records=901325
                Reduce output records=67505
                Spilled Records=1802650
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=285
                CPU time spent (ms)=7740
                Physical memory (bytes) snapshot=721805312
                Virtual memory (bytes) snapshot=7776735232
                Total committed heap usage (bytes)=571998208
                Peak Map Physical memory (bytes)=256131072
                Peak Map Virtual memory (bytes)=2589831168
                Peak Reduce Physical memory (bytes)=212832256
                Peak Reduce Virtual memory (bytes)=2598166528
        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=5462296
        File Output Format Counters
                Bytes Written=717768
2021-03-20 00:26:54,195 INFO streaming.StreamJob: Output directory: output_word
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2$
```

e. Testing if map reduce program in python ran correctly.

```
youth?  5
youthful        29
youthful,       2
youths  4
youtli  1
zanies. 1
zany,   1
zeal    20
zeal!   1
zeal,   7
zeal.   4
zealous 6
zeals,  1
zed!    1
zenith  1
zephyrs 1
zip     1
zir,    1
zir.    1
zo      1
zodiac  1
zodiacs 1
zone,   1
zounds! 1
zounds, 1
zwagger'd       1
}       2
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2$
```