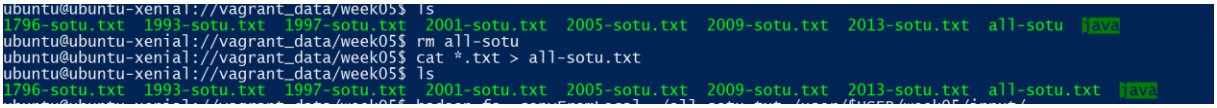
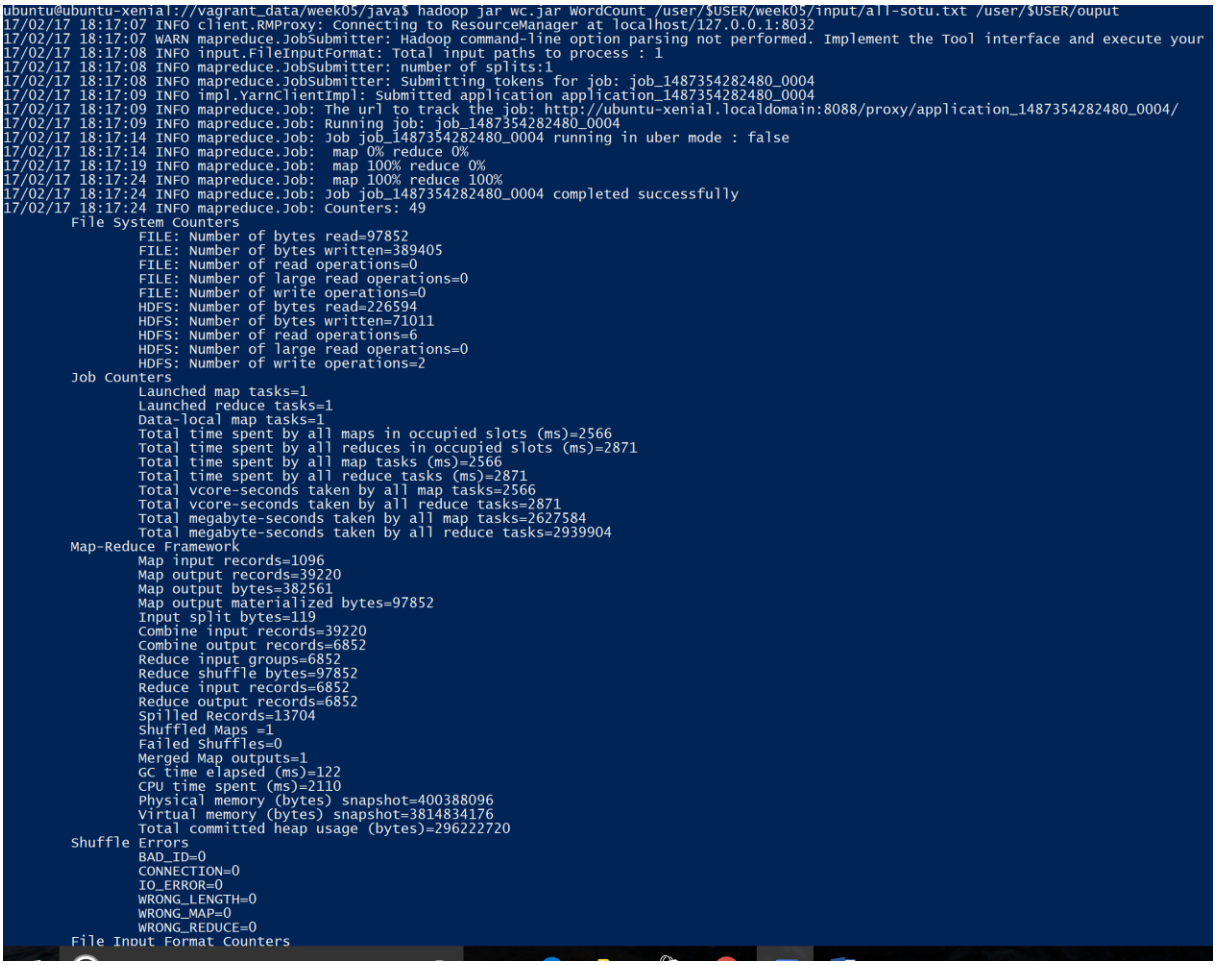
Week-05 I/O – PRADEEP RAJA MOHAN (A20370429)

1. Cat all source files into a single text file



2. Running wordcount v1 jar file in Hadoop



3. Output for wordcount1

```
youthful        1
you'd   1
you'll  1
you've  2
zealous 1
zero    1
zones   1
zones.  1
--      2
—       69
"But    1
"College        1
"Fix-It-First"  1
"I      2
"It     1
"That's 1
"The    1
"There  1
"We     1
"something      1
"the    1
"to     1
ubuntu@ubuntu-xenial://vagrant_data/week05/java$
```

4. Top 10 words for wordcount1 from the output

```
ubuntu@ubuntu-xenial://vagrant_data/pmohan3/itmd521/week-05/output$ hadoop fs -cat /user/$USER/ouput/part-r-00000 | sort -rn -k2 | head -n10
the     1867
to      1433
and     1217
of      1142
a       757
our     657
in      640
that    571
we      560
for     445
ubuntu@ubuntu-xenial://vagrant_data/pmohan3/itmd521/week-05/output$
```

5. Running wordcount v2 for the same input data

```
ubuntu@ubuntu-xenial://vagrant_data/pmohan3/itmd521/week-05/wordcount2$ hadoop jar wc.jar WordCount2 /user/$USER/week05/input/all-sotu.txt /user/$USER/ouput2
17/02/17 18:45:41 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
17/02/17 18:45:41 INFO input.FileInputFormat: Total input paths to process : 1
17/02/17 18:45:41 INFO mapreduce.JobSubmitter: number of splits:1
17/02/17 18:45:42 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1487354282480_0006
17/02/17 18:45:42 INFO impl.YarnClientImpl: Submitted application application_1487354282480_0006
17/02/17 18:45:42 INFO mapreduce.Job: The url to track the job: http://ubuntu-xenial.localdomain:8088/proxy/application_1487354282480_0006/
17/02/17 18:45:42 INFO mapreduce.Job: Running job: job_1487354282480_0006
17/02/17 18:45:47 INFO mapreduce.Job: Job job_1487354282480_0006 running in uber mode : false
17/02/17 18:45:47 INFO mapreduce.Job:  map 0% reduce 0%
17/02/17 18:45:51 INFO mapreduce.Job:  map 100% reduce 0%
17/02/17 18:45:55 INFO mapreduce.Job:  map 100% reduce 100%
17/02/17 18:45:56 INFO mapreduce.Job: Job job_1487354282480_0006 completed successfully
17/02/17 18:45:56 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=97852
                FILE: Number of bytes written=389717
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=226594
                HDFS: Number of bytes written=71011
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        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)=2457
                Total time spent by all reduces in occupied slots (ms)=1764
                Total time spent by all map tasks (ms)=2457
                Total time spent by all reduce tasks (ms)=1764
                Total vcore-seconds taken by all map tasks=2457
                Total vcore-seconds taken by all reduce tasks=1764
                Total megabyte-seconds taken by all map tasks=2515968
                Total megabyte-seconds taken by all reduce tasks=1806336
        Map-Reduce Framework
```

6. Output for wordcout2

```
young   23
younger 6
youngest        1
your    56
yours,  2
yourself        1
yourself,       1
youth   2
youth,  1
youth.  1
youthful        1
you’d   1
you’ll  1
you’ve  2
zealous 1
zero    1
zones   1
zones.  1
--      2
—       69
“But    1
“College        1
“Fix-It-First”  1
“I      2
“It     1
“That’s 1
“The    1
“There  1
“We     1
“something      1
“the    1
“to     1
ubuntu@ubuntu-xenial://vagrant_data/pmohan3/itmd521/week-05/wordcount2$
```

7. Top 10 words for wordcount v2

```
ubuntu@ubuntu-xenial://vagrant_data/pmohan3/itmd521/week-05/output$ hadoop fs -cat /user/$USER/ouput2/part-r-00000 | sort -rn -k2 | head -n10
the     1867
to      1433
and     1217
of      1142
a       757
our     657
in      640
that    571
we      560
for     445
```

8. create jar files for Modified WordCount V1

```
ubuntu@ubuntu-xenial://vagrant_data/pmohan3/itmd521/week-05/modified WordCount1$ hadoop com.sun.tools.javac.Main WordCount.java
ubuntu@ubuntu-xenial://vagrant_data/pmohan3/itmd521/week-05/modified WordCount1$ ls
WordCount.class  WordCount$IntSumReducer.class  WordCount.java  WordCount$TokenizerMapper.class
ubuntu@ubuntu-xenial://vagrant_data/pmohan3/itmd521/week-05/modified WordCount1$ hadoop jar wc.jar WordCount /user/$USER/week05/input/all-sotu.txt /user/$USER/ouput3
Not a valid JAR: /vagrant_data/pmohan3/itmd521/week-05/modified WordCount1/wc.jar
ubuntu@ubuntu-xenial://vagrant_data/pmohan3/itmd521/week-05/modified WordCount1$ jar cf wc.jar WordCount*.class
```

9. Run the modified wordcount v1 jar in Hadoop

```
ubuntu@ubuntu-xenial://vagrant_data/pmohan3/itmd521/week-05/modified WordCount1$ hadoop jar wc.jar WordCount /user/$USER/week05/input/all-sotu.txt /user/$USER/ouput3
17/02/17 19:06:42 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
17/02/17 19:06:42 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner
17/02/17 19:06:43 INFO input.FileInputFormat: Total input paths to process : 1
17/02/17 19:06:43 INFO mapreduce.JobSubmitter: number of splits:1
17/02/17 19:06:43 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1487354282480_0007
17/02/17 19:06:43 INFO impl.YarnClientImpl: Submitted application application_1487354282480_0007
17/02/17 19:06:43 INFO mapreduce.Job: The url to track the job: http://ubuntu-xenial.localdomain:8088/proxy/application_1487354282480_0007/
17/02/17 19:06:43 INFO mapreduce.Job: Running job: job_1487354282480_0007
17/02/17 19:06:47 INFO mapreduce.Job: Job job_1487354282480_0007 running in uber mode : false
17/02/17 19:06:47 INFO mapreduce.Job:  map 0% reduce 0%
17/02/17 19:06:52 INFO mapreduce.Job:  map 100% reduce 0%
17/02/17 19:06:57 INFO mapreduce.Job:  map 100% reduce 100%
17/02/17 19:06:58 INFO mapreduce.Job: Job job_1487354282480_0007 completed successfully
17/02/17 19:06:58 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=17423
                FILE: Number of bytes written=228549
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=226594
                HDFS: Number of bytes written=12638
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        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)=1864
                Total time spent by all reduces in occupied slots (ms)=2219
                Total time spent by all map tasks (ms)=1864
                Total time spent by all reduce tasks (ms)=2219
                Total vcore-seconds taken by all map tasks=1864
                Total vcore-seconds taken by all reduce tasks=2219
                Total megabyte-seconds taken by all map tasks=1908736
                Total megabyte-seconds taken by all reduce tasks=2272256
        Map-Reduce Framework
                Map input records=1096
                Map output records=39220
                Map output bytes=382561
                Map output materialized bytes=17423
                Input split bytes=119
                Combine input records=39220
                Combine output records=1338
                Reduce input groups=1338
                Reduce shuffle bytes=17423
                Reduce input records=1338
                Reduce output records=1338
                Spilled Records=2676
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=98
                CPU time spent (ms)=1380
                Physical memory (bytes) snapshot=397697024
                Virtual memory (bytes) snapshot=3812839424
                Total committed heap usage (bytes)=311427072
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
```

10. Output having word count having more than 4

```
work,   14
work.   10
worked  8
workers 23
workers,        6
workers.        4
working 23
works   7
world   32
world's 11
world,  10
world.  28
worse   5
worthy  4
would   38
year    46
year,   24
year.   11
years   63
years,  30
years.  20
yet     8
you     131
you'll  6
you're  6
you,    12
you.    9
young   23
younger 6
your    56
—       69
ubuntu@ubuntu-xenial://vagrant_data/pmohan3/itmd521/week-05/modified WordCount1$
```

11. Top 10 words for modified wordcount v1

```
ubuntu@ubuntu-xenial://vagrant_data/pmohan3/itmd521/week-05/modified WordCount1$ hadoop fs -cat /user/$USER/ouput31/part-r-00000 | sort -rn -k2 | head -n
10
the     1867
to      1433
and     1217
of      1142
a       757
our     657
in      640
that    571
we      560
for     445
ubuntu@ubuntu-xenial://vagrant_data/pmohan3/itmd521/week-05/modified WordCount1$
```

12. Running modified wordcount2 in Hadoop

```
ubuntu@ubuntu-xenial://vagrant_data/pmohan3/itmd521/week-05/modified WordCount2$ hadoop jar wc.jar WordCount2 /user/$USER/week05/input/all-sotu.txt /user/$USER/ouput4 -skip /user/$USER/week05/input/pattern.txt
17/02/18 01:29:36 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
17/02/18 01:29:37 INFO input.FileInputFormat: Total input paths to process : 1
17/02/18 01:29:37 INFO mapreduce.JobSubmitter: number of splits:1
17/02/18 01:29:37 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1487354282480_0010
17/02/18 01:29:37 INFO impl.YarnClientImpl: Submitted application application_1487354282480_0010
17/02/18 01:29:37 INFO mapreduce.Job: The url to track the job: http://ubuntu-xenial.localdomain:8088/proxy/application_1487354282480_0010/
17/02/18 01:29:37 INFO mapreduce.Job: Running job: job_1487354282480_0010
17/02/18 01:29:41 INFO mapreduce.Job: Job job_1487354282480_0010 running in uber mode : false
17/02/18 01:29:41 INFO mapreduce.Job:  map 0% reduce 0%
17/02/18 01:29:46 INFO mapreduce.Job:  map 100% reduce 0%
17/02/18 01:29:50 INFO mapreduce.Job:  map 100% reduce 100%
17/02/18 01:29:51 INFO mapreduce.Job: Job job_1487354282480_0010 completed successfully
17/02/18 01:29:51 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=89691
                FILE: Number of bytes written=375411
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=226594
                HDFS: Number of bytes written=63581
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        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)=2182
                Total time spent by all reduces in occupied slots (ms)=1815
                Total time spent by all map tasks (ms)=2182
                Total time spent by all reduce tasks (ms)=1815
                Total vcore-seconds taken by all map tasks=2182
                Total vcore-seconds taken by all reduce tasks=1815
                Total megabyte-seconds taken by all map tasks=2234368
                Total megabyte-seconds taken by all reduce tasks=1858560
        Map-Reduce Framework
                Map input records=1096
                Map output records=33826
                Map output bytes=324293
                Map output materialized bytes=89691
                Input split bytes=119
                Combine input records=33826
                Combine output records=6660
                Reduce input groups=6660
                Reduce shuffle bytes=89691
                Reduce input records=6660
                Reduce output records=6660
                Spilled Records=13320
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=104
                CPU time spent (ms)=1770
                Physical memory (bytes) snapshot=433438720
                Virtual memory (bytes) snapshot=3820085248
                Total committed heap usage (bytes)=301465600
```

13. Output for running modifiedwordcount v2

```
yet     18
yeto    1
yetternational  1
yield   3
yogi    1
york    4
you     162
youcrease       1
youll   6
young   23
younger 6
youngest        1
youngstown      1
younight        2
your    51
youre   7
yourput 1
yourquiry       1
yours   2
yourself        2
yoursurance     1
yourtention     1
yourvestments   1
yourvitation    1
youth   4
youthful        1
youve   3
you'd   1
you'll  1
you've  2
zarfos  3
zarqawi 1
zealoustention  1
zerolerance     1
zones   2
ubuntu@ubuntu-xenial://vagrant_data/pmohan3/itmd521/week-05/modified WordCount2$
```

14. Top 10 words

```
ubuntu@ubuntu-xenial://vagrant_data/pmohan3/itmd521/week-05/modified WordCount2$ hadoop fs -cat /user/$USER/output4-4/part-r-00000 | sort -rn -k2 | head -n10
the     1887
and     1367
a       760
we      738
our     688
that    594
is      381
will    378
i       307
this    303
```

15. Moving output file for analysis from HDFS

```
ubuntu@ubuntu-xenial://vagrant_data/pmohan3/itmd521/week-05$ cd output/
ubuntu@ubuntu-xenial://vagrant_data/pmohan3/itmd521/week-05/output$ hadoop fs -cat /user/$USER/ouput/part-r-00000>wordcountv1.txt
ubuntu@ubuntu-xenial://vagrant_data/pmohan3/itmd521/week-05/output$ hadoop fs -cat /user/$USER/ouput2/part-r-00000>wordcountv2.txt
ubuntu@ubuntu-xenial://vagrant_data/pmohan3/itmd521/week-05/output$ hadoop fs -cat /user/$USER/ouput31/part-r-00000>modifiedwordcountv1.txt
```

ANALYSIS:

1. Number of words for running wordcountv1 : 6852
2. Number of words for running wordcountv2 : 6852
3. Number of words for running Modified wordcountv1 : 1089
4. Number of words for running Modified wordcountv2 : 5437