분산시스템 과제 #4

박종현, 2025-04-17
공과대학 컴퓨터정보통신공학과

과제 목표
1. 2-8 자료 : 7, 8, 9, 14, 15쪽
2. 2-9 자료 : 11, 12, 19, 20쪽
3. 2-10 자료 : 8, 9, 17, 18쪽

[Ch 2-8 p.7-9]

```
!       2
!!home   4
!!home\t        1
!!ht    1
!action 2
!coaches        1
!width=60       5
!"      1
"       487
""      10
""bonus 1
""i     1
""it's  1
""marry 1
""mirror        1
""missin'       1
""moon" 1
""ok    1
""polish"-similarly     1
""rumor 1
""the   1
""we    2
""you're        1
"$50    1
"'      1
"'''    1
--More--
```

[Ch 2-8 p.14-15]


```
hduser@DESKTOP-635D6IV:~/hadoop_test/citationcount$ javac -classpath $(hadoop classpath) -d bin src/*.java
Note: src/CountCitation.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
hduser@DESKTOP-635D6IV:~/hadoop_test/citationcount$ jar -cvf citationcount_auth.jar -C bin .
added manifest
adding: CountCitation.class(in = 1768) (out= 878)(deflated 50%)
adding: CountCitation$Reduce.class(in = 1639) (out= 688)(deflated 58%)
adding: CountCitation$Map.class(in = 1424) (out= 546)(deflated 61%)
hduser@DESKTOP-635D6IV:~/hadoop_test/citationcount$ hadoop jar citationcount_auth.jar CountCitation input/2M.SRCID.DSTID
 cc_out_auth1
2025-04-17 01:38:14,604 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2025-04-17 01:38:15,101 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn-staging/
hduser/.staging/job_1744821131799_0003
2025-04-17 01:38:15,375 INFO input.FileInputFormat: Total input files to process : 1
2025-04-17 01:38:16,294 INFO mapreduce.JobSubmitter: number of splits:2
2025-04-17 01:38:16,859 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744821131799_0003
2025-04-17 01:38:16,859 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-04-17 01:38:17,052 INFO conf.Configuration: resource-types.xml not found
2025-04-17 01:38:17,052 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-04-17 01:38:17,159 INFO impl.YarnClientImpl: Submitted application application_1744821131799_0003
2025-04-17 01:38:17,221 INFO mapreduce.Job: The url to track the job: http://DESKTOP-635D6IV.:50000/proxy/application_17
44821131799_0003/
2025-04-17 01:38:17,221 INFO mapreduce.Job: Running job: job_1744821131799_0003
2025-04-17 01:38:27,389 INFO mapreduce.Job: Job job_1744821131799_0003 running in uber mode : false
2025-04-17 01:38:27,391 INFO mapreduce.Job:  map 0% reduce 0%
2025-04-17 01:38:37,568 INFO mapreduce.Job:  map 50% reduce 0%
```



```
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=631
                CPU time spent (ms)=32580
                Physical memory (bytes) snapshot=1301839872
                Virtual memory (bytes) snapshot=7700021248
                Total committed heap usage (bytes)=1278214144
                Peak Map Physical memory (bytes)=616513536
                Peak Map Virtual memory (bytes)=2570010624
                Peak Reduce Physical memory (bytes)=335585280
                Peak Reduce Virtual memory (bytes)=2567725056
        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=151645466
        File Output Format Counters
                Bytes Written=12476697
hduser@DESKTOP-635D6IV:~/hadoop_test/citationcount$ hdfs dfs -ls cc_out_auth1
Found 2 items
-rw-r--r--   1 hduser supergroup          0 2025-04-17 01:38 cc_out_auth1/_SUCCESS
-rw-r--r--   1 hduser supergroup   12476697 2025-04-17 01:38 cc_out_auth1/part-r-00000
hduser@DESKTOP-635D6IV:~/hadoop_test/citationcount$
```

```
10000001    2
10000010    2
10000017    1
10000030    1
1000007 2
1000016 3
10000176    8
10000218    3
10000230    4
10000244    2
10000288    1
1000030 36
10000323    4
1000035 24
10000383    3
1000044 13
10000462    6
10000495    2
10000498    2
100005  112
10000514    1
10000530    1
10000548    8
10000591    4
10000649    1
10000705    4
--More--
```

```java
*/
public class CountCitation {

  public static class Map extends Mapper<Text, Text, Text, IntWritable> {
      private final static IntWritable one = new IntWritable(1);

      public void map(Text key, Text value, Context context) throws IOException, InterruptedException {

          context.write(key, one);

      }
  }

  public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable> {

      public void reduce(Text key, Iterable<IntWritable> values, Context context)
        throws IOException, InterruptedException {
        int sum = 0;
        for (IntWritable val : values) {
            sum += val.get();
        }
        context.write(key, new IntWritable(sum));
      }
  }
}
src/CountCitation.java [+]                                          22,26        31%
-- INSERT --
```
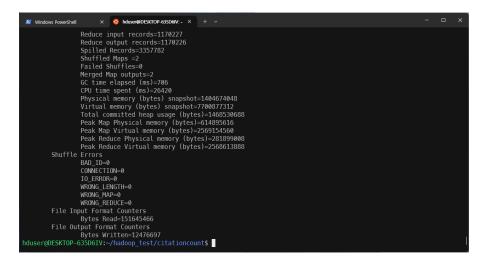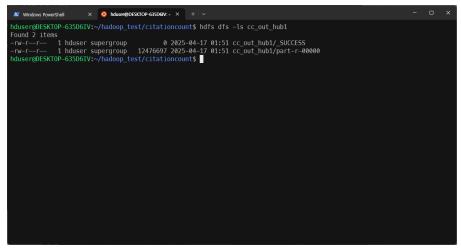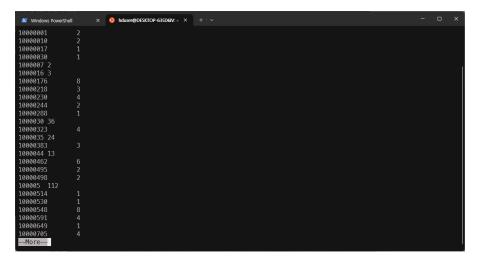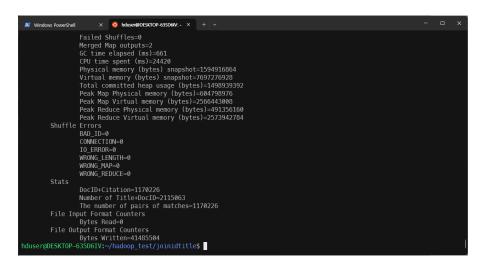
```
        at javax.security.auth.Subject.doAs(Subject.java:422)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1953)
        at org.apache.hadoop.mapreduce.Job.submit(Job.java:1674)
        at org.apache.hadoop.mapreduce.Job.waitForCompletion(Job.java:1695)
        at CountCitation.main(CountCitation.java:63)
        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:330)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:245)
hduser@DESKTOP-635D6IV:~/hadoop_test/citationcount$ hadoop jar citationcount_hub.jar CountCitation input/2M.SRCID.DSTID
cc_out_hub1
2025-04-17 01:51:05,085 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2025-04-17 01:51:05,500 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/
hduser/.staging/job_1744821131799_0004
2025-04-17 01:51:05,821 INFO input.FileInputFormat: Total input files to process : 1
2025-04-17 01:51:06,734 INFO mapreduce.JobSubmitter: number of splits:2
2025-04-17 01:51:07,310 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744821131799_0004
2025-04-17 01:51:07,310 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-04-17 01:51:07,501 INFO conf.Configuration: resource-types.xml not found
2025-04-17 01:51:07,501 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-04-17 01:51:07,598 INFO impl.YarnClientImpl: Submitted application application_1744821131799_0004
2025-04-17 01:51:07,655 INFO mapreduce.Job: The url to track the job: http://DESKTOP-635D6IV.:50000/proxy/application_17
44821131799_0004/
2025-04-17 01:51:07,656 INFO mapreduce.Job: Running job: job_1744821131799_0004
```
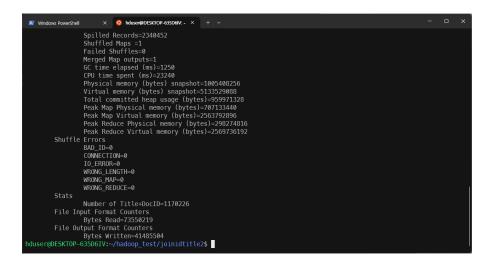
```
              Reduce input records=1170227
              Reduce output records=1170226
              Spilled Records=3357782
              Shuffled Maps =2
              Failed Shuffles=0
              Merged Map outputs=2
              GC time elapsed (ms)=706
              CPU time spent (ms)=26420
              Physical memory (bytes) snapshot=1404674048
              Virtual memory (bytes) snapshot=7700877312
              Total committed heap usage (bytes)=1468530688
              Peak Map Physical memory (bytes)=614895616
              Peak Map Virtual memory (bytes)=2569154560
              Peak Reduce Physical memory (bytes)=281899008
              Peak Reduce Virtual memory (bytes)=2568613888
      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=151645466
      File Output Format Counters
              Bytes Written=12476697
hduser@DESKTOP-635D6IV:~/hadoop_test/citationcount$
```

```
hduser@DESKTOP-635D6IV:~/hadoop_test/citationcount$ hdfs dfs -ls cc_out_hub1
Found 2 items
-rw-r--r--   1 hduser supergroup          0 2025-04-17 01:51 cc_out_hub1/_SUCCESS
-rw-r--r--   1 hduser supergroup   12476697 2025-04-17 01:51 cc_out_hub1/part-r-00000
hduser@DESKTOP-635D6IV:~/hadoop_test/citationcount$
```

```
10000001        2
10000010        2
10000017        1
10000030        1
1000007 2
1000016 3
10000176        8
10000218        3
10000230        4
10000244        2
10000288        1
1000030 36
10000323        4
1000035 24
10000383        3
1000044 13
10000462        6
10000495        2
10000498        2
100005  112
10000514        1
10000530        1
10000548        8
10000591        4
10000649        1
10000705        4
--More--
```

[Ch 2-9 p.11-12]


```
hduser@DESKTOP-635D6IV:~/hadoop_test/joinidtitle$ hadoop jar joinidtitle.jar JoinIDTitle input/2M.TITLE.ID cc_out_auth1/
part-r-00000 join_out1
2025-04-17 02:07:53,330 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2025-04-17 02:07:53,681 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement
the Tool interface and execute your application with ToolRunner to remedy this.
2025-04-17 02:07:53,718 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/
hduser/.staging/job_1744821131799_0011
2025-04-17 02:07:53,970 INFO input.FileInputFormat: Total input files to process : 1
2025-04-17 02:07:53,996 INFO input.FileInputFormat: Total input files to process : 1
2025-04-17 02:07:54,466 INFO mapreduce.JobSubmitter: number of splits:2
2025-04-17 02:07:54,588 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744821131799_0011
2025-04-17 02:07:54,588 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-04-17 02:07:54,769 INFO conf.Configuration: resource-types.xml not found
2025-04-17 02:07:54,769 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'
2025-04-17 02:07:54,881 INFO impl.YarnClientImpl: Submitted application application_1744821131799_0011
2025-04-17 02:07:54,920 INFO mapreduce.Job: The url to track the job: http://DESKTOP-635D6IV.:50000/proxy/application_17
44821131799_0011/
2025-04-17 02:07:54,920 INFO mapreduce.Job: Running job: job_1744821131799_0011
2025-04-17 02:08:01,026 INFO mapreduce.Job: Job job_1744821131799_0011 running in uber mode : false
2025-04-17 02:08:01,027 INFO mapreduce.Job:  map 0% reduce 0%
2025-04-17 02:08:09,130 INFO mapreduce.Job:  map 50% reduce 0%
2025-04-17 02:08:15,174 INFO mapreduce.Job:  map 100% reduce 0%
2025-04-17 02:08:21,208 INFO mapreduce.Job:  map 100% reduce 100%
2025-04-17 02:08:22,229 INFO mapreduce.Job: Job job_1744821131799_0011 completed successfully
2025-04-17 02:08:22,359 INFO mapreduce.Job: Counters: 58
        File System Counters
                FILE: Number of bytes read=181179111
```



```
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=661
                CPU time spent (ms)=24420
                Physical memory (bytes) snapshot=1594916864
                Virtual memory (bytes) snapshot=7697276928
                Total committed heap usage (bytes)=1498939392
                Peak Map Physical memory (bytes)=604798976
                Peak Map Virtual memory (bytes)=2566443008
                Peak Reduce Physical memory (bytes)=491356160
                Peak Reduce Virtual memory (bytes)=2573942784
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        Stats
                DocID+Citation=1170226
                Number of Title+DocID=2115063
                The number of pairs of matches=1170226
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=41485504
hduser@DESKTOP-635D6IV:~/hadoop_test/joinidtitle$
```



```
hduser@DESKTOP-635D6IV:~/hadoop_test/joinidtitle$ hdfs dfs -ls join_out1
Found 2 items
-rw-r--r--   1 hduser supergroup          0 2025-04-17 02:08 join_out1/_SUCCESS
-rw-r--r--   1 hduser supergroup   41485504 2025-04-17 02:08 join_out1/part-r-00000
hduser@DESKTOP-635D6IV:~/hadoop_test/joinidtitle$ hdfs dfs -cat join_out1/part-r-00000 | head -5
10000001        Juan que reía   2
10000010        Golden bear oil 2
10000017        La Isla de los dibujos  1
10000030        Los Hombres sólo piensan en eso 1
1000007 Ettercap        2
cat: Unable to write to output stream.
hduser@DESKTOP-635D6IV:~/hadoop_test/joinidtitle$
```

[Ch 2-9 p.19-20]

[Ch 2-10 p.8-9]

[Ch 2-10 p.17-18]

```
                Peak Reduce Physical memory (bytes)=384638976
                Peak Reduce Virtual memory (bytes)=2573369344
        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=9735683
        File Output Format Counters
                Bytes Written=6556893
hduser@DESKTOP-635D6IV:~/hadoop_test/invertedindex2$ hdfs dfs -ls ii2_out1
Found 2 items
-rw-r--r--   1 hduser supergroup          0 2025-04-17 02:21 ii2_out1/_SUCCESS
-rw-r--r--   1 hduser supergroup    6556893 2025-04-17 02:21 ii2_out1/part-r-00000
hduser@DESKTOP-635D6IV:~/hadoop_test/invertedindex2$ hdfs dfs -cat ii2_out1/part-r-00000 | head -5
0       2185,39406,46953,56557,59739,105290,107570,107600,107603,110266,111115,112036,114463,114604,114935,115462,115647
,116152,116421,117317,117402,119358,119531,119565,119591,119971,120417,122885,123128,126671,127550,127824,128976,130031,
130269,131301,131343,134128,134896,135674,137366,139104,139765,150480,150962,213900,231291,259374,259640,259953,260233,7
36861,821165,846960,863472,873177,922939,953399,991203,1083631,1157839,1201310,1398193,1429917,1544374,1669676,1722590,1
851995,1888165,1932784,1999660,2030470,2333061,2439703,2447166,2522680,2554996,3038004,3120913,3159417,3174112,3414795,3
509072,3587831,3708222,3869884,4037764,4064809,4075106,4095197,4530540,4606153,4890927,5042786,5171379,5343488,5632253,5
655279,5662644,5721663,5876585,5901876,6255582,6262417,6554636,6554748,6704337,6717145,6816176,6916196,6930599,6937695,7
081129,7945204,8276444,8678146,8773592,8857587,8923503,9007073,9173011,9263305,9538068,9626096,9752106,9885025,10801886,
11979629,12072140,12072986,12092675,12242363,13291495,13592764,13737413,13753904,13957506,14160104,14205946,14431940,148
```