Read me

Question 1:

Please copy the Jar file from project folder into CS6360 server.

Steps to Execute:

1. If files are readily available, directly go to step 2 otherwise execute following command for download files

hadoop jar combiner-0.2-SNAPSHOT.jar bigdata.combiner.DownloadBigData /user/kpw150030/A5_Files

This program accepts only one location - a non existing directory where files would be downloaded.

2. Execute Following command in order to execute word count example with in mapper combiner implementation.

start=$SECONDS;hadoop jar combiner-0.2-SNAPSHOT.jar bigdata.combiner.WordCountWithInMapperCombiner /user/kpw150030/A5_Files /user/kpw150030/A5_IN_OP;duration=$((SECONDS - start));echo 'The running time of WordCount in Mapreduce is:';echo $duration

This Program accepts two inputs - first is the location where input files are stored and second one is for writing output.

As we can see in following output, the program gets executed within 22 seconds.

3. Execute following command for executing class in in which External Combiner is defined as the reducer class it self.

start=$SECONDS;hadoop jar combiner-0.2-SNAPSHOT.jar bigdata.combiner.WordCountExternalCombiner /user/kpw150030/A5_Files /user/kpw150030/A5_EXT_OP;duration=$((SECONDS - start));echo 'The running time of WordCount in Mapreduce is:';echo $duration

As we can see, this takes 23 seconds, one second extra as compared to the in mapper combiner. Also, we can see the mapper output size is same as combiner output size. This definitely proves how in-mapper combiner is more efficient than an external combiner.

Map-Reduce Framework (With external combiner)

Map output records=1815930
Map output bytes=17684772
Combine output records=285614
Reduce input records=285614

Map-Reduce Framework (In mapper combiner

Map output records=285614
Map output bytes=3585712
Combine input records=0
Combine output records=0
Reduce input records=285614

Question 2:

Steps to Execute Program

Please execute following command

hadoop jar combiner-0.2-SNAPSHOT.jar bigdata.combiner.CoOccuranceWithStripes /user/kpw150030/B4_Files /user/kpw150030/Stripe_COOCurrance /user/kpw150030/stopwords_en.txt 5 2;

Command Explanation : This program accepts 5 parameters

[INPUT_DIR] [OUTPUT_DIR] [STOP_WORDS_FILE] [MIN_WORD_LENGTH] [NUM_NEIGHBOURS]

1. Input file
2. Output file location
3. StopWords File path
4. Minimum WordLength
5. num neighbors - Please provide value greater than number 2

Please find stopwords file and sample output in docs folder next to this ReadMe file.

Detailed Execution Log of Question 1:

{cs6360:~} start=$SECONDS;hadoop jar combiner-0.2-SNAPSHOT.jar bigdata.combiner.WordCountWithInMapperCombiner /user/kpw150030/A5_Files /user/kpw150030/A5_IN1_OP;duration=$((SECONDS - start));echo 'The running time of WordCount in Mapreduce is:';echo $duration

16/02/17 18:30:01 INFO client.RMProxy: Connecting to ResourceManager at cshadoop1.utdallas.edu/10.176.92.71:8032
16/02/17 18:30:02 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
16/02/17 18:30:02 INFO input.FileInputFormat: Total input paths to process : 12
16/02/17 18:30:02 INFO mapreduce.JobSubmitter: number of splits:12

16/02/17 18:30:02 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
16/02/17 18:30:03 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1455750748565_0056
16/02/17 18:30:03 INFO impl.YarnClientImpl: Submitted application application_1455750748565_0056
16/02/17 18:30:03 INFO mapreduce.Job: The url to track the job: http://cshadoop1.utdallas.edu:8088/proxy/application_1455750748565_0056/
16/02/17 18:30:03 INFO mapreduce.Job: Running job: job_1455750748565_0056
16/02/17 18:30:08 INFO mapreduce.Job: Job job_1455750748565_0056 running in uber mode : false
16/02/17 18:30:08 INFO mapreduce.Job: map 0% reduce 0%
16/02/17 18:30:12 INFO mapreduce.Job: map 8% reduce 0%
16/02/17 18:30:13 INFO mapreduce.Job: map 17% reduce 0%
16/02/17 18:30:14 INFO mapreduce.Job: map 42% reduce 0%
16/02/17 18:30:16 INFO mapreduce.Job: map 75% reduce 0%
16/02/17 18:30:17 INFO mapreduce.Job: map 92% reduce 0%
16/02/17 18:30:18 INFO mapreduce.Job: map 100% reduce 0%
16/02/17 18:30:20 INFO mapreduce.Job: map 100% reduce 100%
16/02/17 18:30:20 INFO mapreduce.Job: Job job_1455750748565_0056 completed successfully
16/02/17 18:30:20 INFO mapreduce.Job: Counters: 50
    File System Counters
        FILE: Number of bytes read=4156946
        FILE: Number of bytes written=9525661
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=7007071
        HDFS: Number of bytes written=1084900
        HDFS: Number of read operations=39
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
    Job Counters
        Launched map tasks=12
        Launched reduce tasks=1
        Data-local map tasks=11
        Rack-local map tasks=1
        Total time spent by all maps in occupied slots (ms)=62635
        Total time spent by all reduces in occupied slots (ms)=4197
        Total time spent by all map tasks (ms)=62635
        Total time spent by all reduce tasks (ms)=4197
        Total vcore-seconds taken by all map tasks=62635
        Total vcore-seconds taken by all reduce tasks=4197
        Total megabyte-seconds taken by all map tasks=64138240
        Total megabyte-seconds taken by all reduce tasks=4297728
    Map-Reduce Framework
        Map input records=209490
        Map output records=285614
        Map output bytes=3585712
        Map output materialized bytes=4157012
        Input split bytes=1400

```
        Combine input records=0
        Combine output records=0
        Reduce input groups=98000
        Reduce shuffle bytes=4157012
        Reduce input records=285614
        Reduce output records=98000
        Spilled Records=571228
        Shuffled Maps =12
        Failed Shuffles=0
        Merged Map outputs=12
        GC time elapsed (ms)=659
        CPU time spent (ms)=21740
        Physical memory (bytes) snapshot=4619235328
        Virtual memory (bytes) snapshot=16024215552
        Total committed heap usage (bytes)=6577192960
    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=7005671
    File Output Format Counters
        Bytes Written=1084900
The running time of WordCount in Mapreduce is:
22
{cs6360:~}
{cs6360:~} start=$SECONDS;hadoop jar combiner-0.2-SNAPSHOT.jar bigdata.combiner.WordCountExternalCombiner
/user/kpw150030/A5_Files /user/kpw150030/A5_EXT1_OP;duration=$(( SECONDS - start ));echo 'The running time of
WordCount in Mapreduce is:';echo $duration

16/02/17 18:30:24 INFO client.RMProxy: Connecting to ResourceManager at
cshadoop1.utdallas.edu/10.176.92.71:8032
16/02/17 18:30:25 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed.
Implement the Tool interface and execute your application with ToolRunner to remedy this.
16/02/17 18:30:25 INFO input.FileInputFormat: Total input paths to process : 12
16/02/17 18:30:25 INFO mapreduce.JobSubmitter: number of splits:12
16/02/17 18:30:26 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use
mapreduce.jobtracker.address
16/02/17 18:30:26 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1455750748565_0058
16/02/17 18:30:26 INFO impl.YarnClientImpl: Submitted application application_1455750748565_0058
16/02/17 18:30:26 INFO mapreduce.Job: The url to track the job:
http://cshadoop1.utdallas.edu:8088/proxy/application_1455750748565_0058/
16/02/17 18:30:26 INFO mapreduce.Job: Running job: job_1455750748565_0058
16/02/17 18:30:32 INFO mapreduce.Job: Job job_1455750748565_0058 running in uber mode : false
16/02/17 18:30:32 INFO mapreduce.Job:  map 0% reduce 0%
```

```
16/02/17 18:30:37 INFO mapreduce.Job:  map 17% reduce 0%
16/02/17 18:30:38 INFO mapreduce.Job:  map 33% reduce 0%
16/02/17 18:30:40 INFO mapreduce.Job:  map 67% reduce 0%
16/02/17 18:30:41 INFO mapreduce.Job:  map 92% reduce 0%
16/02/17 18:30:42 INFO mapreduce.Job:  map 100% reduce 0%
16/02/17 18:30:43 INFO mapreduce.Job:  map 100% reduce 100%
16/02/17 18:30:44 INFO mapreduce.Job: Job job_1455750748565_0058 completed successfully
16/02/17 18:30:44 INFO mapreduce.Job: Counters: 50
    File System Counters
        FILE: Number of bytes read=4156946
        FILE: Number of bytes written=9528027
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=7007071
        HDFS: Number of bytes written=1084900
        HDFS: Number of read operations=39
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
    Job Counters
        Launched map tasks=12
        Launched reduce tasks=1
        Data-local map tasks=11
        Rack-local map tasks=1
        Total time spent by all maps in occupied slots (ms)=68703
        Total time spent by all reduces in occupied slots (ms)=3821
        Total time spent by all map tasks (ms)=68703
        Total time spent by all reduce tasks (ms)=3821
        Total vcore-seconds taken by all map tasks=68703
        Total vcore-seconds taken by all reduce tasks=3821
        Total megabyte-seconds taken by all map tasks=70351872
        Total megabyte-seconds taken by all reduce tasks=3912704
    Map-Reduce Framework
        Map input records=209490
        Map output records=1815930
        Map output bytes=17684772
        Map output materialized bytes=4157012
        Input split bytes=1400
        Combine input records=1815930
        Combine output records=285614
        Reduce input groups=98000
        Reduce shuffle bytes=4157012
        Reduce input records=285614
        Reduce output records=98000
        Spilled Records=571228
        Shuffled Maps =12
        Failed Shuffles=0
        Merged Map outputs=12
```

GC time elapsed (ms)=237
CPU time spent (ms)=27780
Physical memory (bytes) snapshot=4366630912
Virtual memory (bytes) snapshot=15974055936
Total committed heap usage (bytes)=6563561472
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=7005671
File Output Format Counters
Bytes Written=1084900
The running time of WordCount in Mapreduce is:
23

Detailed Execution log of Question 2:

```
{cs6360:~} hadoop jar combiner-0.2-SNAPSHOT.jar bigdata.combiner.CoOccuranceWith
Stripes /user/kpw150030/B4_Files  /user/kpw150030/Stripe_COOCurrance  /user/kpw1
50030/stopwords_en.txt 5 2;

16/02/19 09:34:25 INFO client.RMProxy: Connecting to ResourceManager at cshadoop
1.utdallas.edu/10.176.92.71:8032
16/02/19 09:34:25 INFO mapreduce.JobSubmissionFiles: Permissions on staging dire
ctory /tmp/hadoop-yarn/staging/kpw150030/.staging are incorrect: rwxrwxrwx. Fixi
ng permissions to correct value rwx------
16/02/19 09:34:26 INFO input.FileInputFormat: Total input paths to process : 6
16/02/19 09:34:26 INFO mapreduce.JobSubmitter: number of splits:6
16/02/19 09:34:26 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_14
55750748565_1962
16/02/19 09:34:26 INFO impl.YarnClientImpl: Submitted application application_14
55750748565_1962
16/02/19 09:34:27 INFO mapreduce.Job: The url to track the job: http://cshadoop1
.utdallas.edu:8088/proxy/application_1455750748565_1962/
16/02/19 09:34:27 INFO mapreduce.Job: Running job: job_1455750748565_1962
16/02/19 09:34:31 INFO mapreduce.Job: Job job_1455750748565_1962 running in uber
mode : false
16/02/19 09:34:31 INFO mapreduce.Job:  map 0% reduce 0%
16/02/19 09:34:36 INFO mapreduce.Job:  map 50% reduce 0%
16/02/19 09:34:37 INFO mapreduce.Job:  map 100% reduce 0%
16/02/19 09:34:41 INFO mapreduce.Job:  map 100% reduce 100%
16/02/19 09:34:42 INFO mapreduce.Job: Job job_1455750748565_1962 completed successfully
16/02/19 09:34:42 INFO mapreduce.Job: Counters: 50
        File System Counters
```

FILE: Number of bytes read=1349751
FILE: Number of bytes written=3362909
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=5382705
HDFS: Number of bytes written=377105
HDFS: Number of read operations=21
HDFS: Number of large read operations=0
HDFS: Number of write operations=2
Job Counters
Launched map tasks=6
Launched reduce tasks=1
Data-local map tasks=2
Rack-local map tasks=4
Total time spent by all maps in occupied slots (ms)=19804
Total time spent by all reduces in occupied slots (ms)=3230
Total time spent by all map tasks (ms)=19804
Total time spent by all reduce tasks (ms)=3230
Total vcore-seconds taken by all map tasks=19804
Total vcore-seconds taken by all reduce tasks=3230
Total megabyte-seconds taken by all map tasks=20279296
Total megabyte-seconds taken by all reduce tasks=3307520
Map-Reduce Framework
Map input records=104745
Map output records=37059
Map output bytes=1275627
Map output materialized bytes=1349781
Input split bytes=688
Combine input records=0
Combine output records=0
Reduce input groups=4497
Reduce shuffle bytes=1349781
Reduce input records=37059
Reduce output records=4497
Spilled Records=74118
Shuffled Maps =6
Failed Shuffles=0
Merged Map outputs=6
GC time elapsed (ms)=439
CPU time spent (ms)=14810
Physical memory (bytes) snapshot=2731720704
Virtual memory (bytes) snapshot=8619110400
Total committed heap usage (bytes)=3554672640
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=5382017
File Output Format Counters
Bytes Written=377105