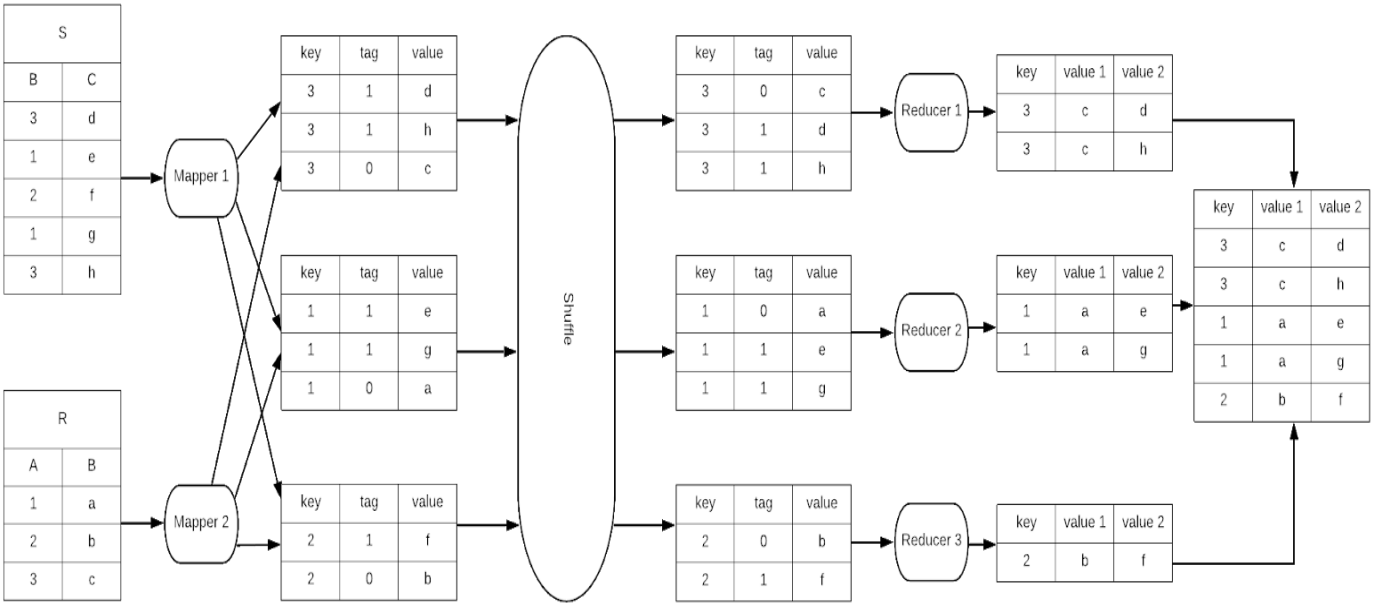
Tommaso Carraro
student number: 015147592

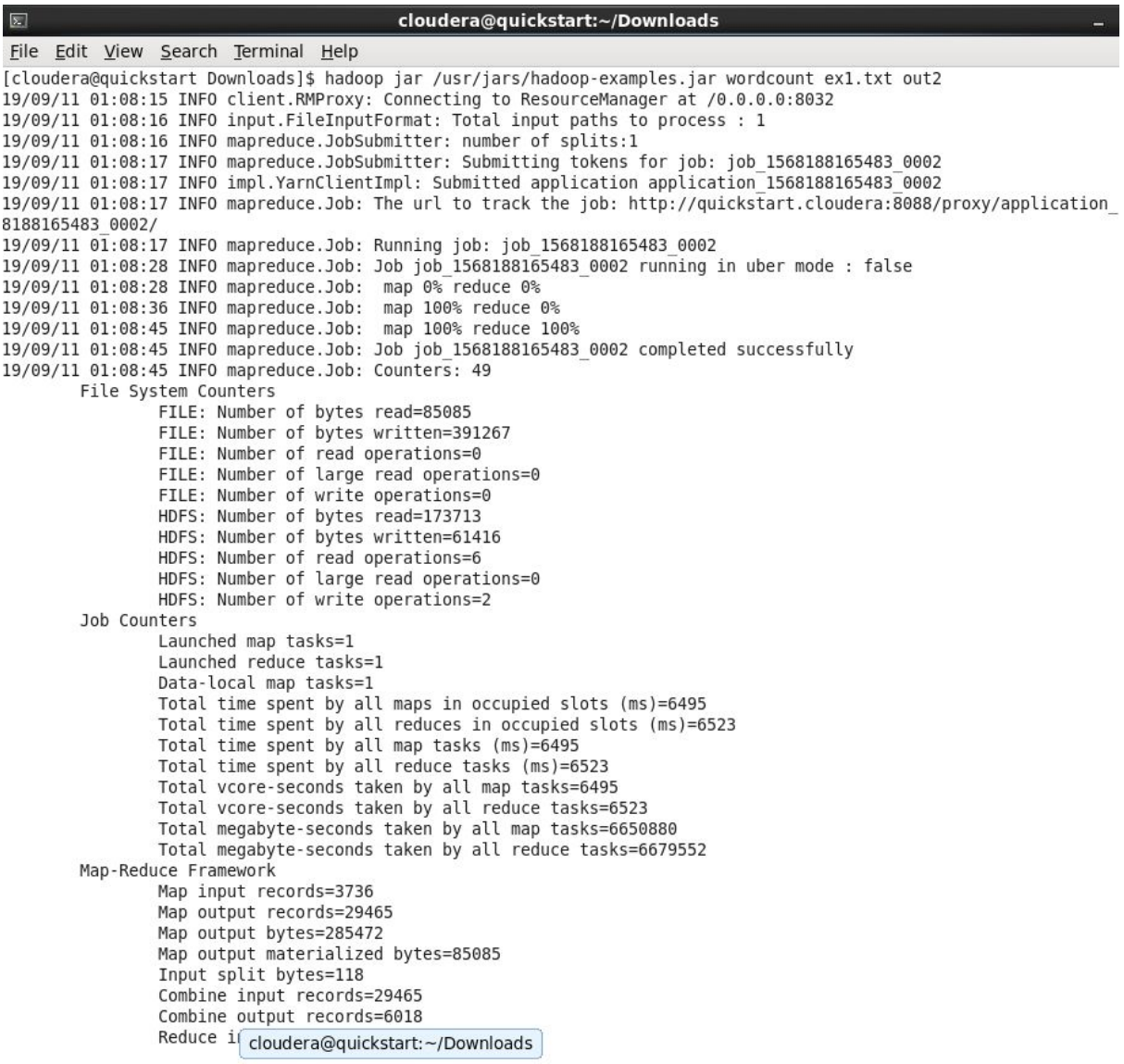
Big Data Management: Homework 2

1.a) In the Map phase there is a mapper for each table. Each mapper takes a table in input and produces partitions based on the key of the table. While it is doing these partitions it also adds a table tag to each row. The mapper that is working on the smaller table adds a table tag of 0 to each row, while the mapper that is working on the larger table adds a table tag of 1 to each row. The output of the Map phase is a set of partitions based on the key value with the adding of a table tag. In the Shuffle phase each partition is sorted based on the value of the table tag. Finally, in the Reduce phase there is a reducer for each partition of the Shuffle phase's output. Each reducer generates the join output for the partition it has received in input and once every reducer has finished its job all the rows are merged together to compose the final join output. A small example is provided:



1.b) In this MapReduce procedure there is only the Map phase because all the work is done by the mappers. This procedure is usually used when there is a really small table that can be stored in the memory of the mapper and a larger table that needs to be splitted to permit the computation. In the Map phase there is a mapper for each split of the larger table. Each mapper stores a copy of the small table in its cache and works only on one split of the larger table. The work of the mapper is to iterate over the split of the larger table performing an hash join with the rows of the smaller table stored in its cache. The output of each mapper is a part of the final join. It is possible to obtain the final output by merging the output of each mapper in one single table. I avoid to make an example of this MapReduce procedure because it is really simple to understand.

2) The world “Cheshire” occurrs 6 times. These are the screens from the virtual machine.



```
Map-Reduce Framework
        Map input records=3736
        Map output records=29465
        Map output bytes=285472
        Map output materialized bytes=85085
        Input split bytes=118
        Combine input records=29465
        Combine output records=6018
        Reduce input groups=6018
        Reduce shuffle bytes=85085
        Reduce input records=6018
        Reduce output records=6018
        Spilled Records=12036
        Shuffled Maps =1
        Failed Shuffles=0
        Merged Map outputs=1
        GC time elapsed (ms)=160
        CPU time spent (ms)=2800
        Physical memory (bytes) snapshot=347586560
        Virtual memory (bytes) snapshot=3007348736
        Total committed heap usage (bytes)=226365440
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=173595
File Output Format Counters
        Bytes Written=61416
```

3) The median world length is 4. These are the screens from the virtual machine.

```
[cloudera@quickstart Downloads]$ hadoop jar /usr/jars/hadoop-examples.jar wordmedian words.txt out3
19/09/11 01:18:30 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
19/09/11 01:18:30 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Implement the Tool i
nterface and execute your application with ToolRunner to remedy this.
19/09/11 01:18:31 INFO input.FileInputFormat: Total input paths to process : 1
19/09/11 01:18:31 INFO mapreduce.JobSubmitter: number of splits:1
19/09/11 01:18:31 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1568188165483_0003
19/09/11 01:18:32 INFO impl.YarnClientImpl: Submitted application application_1568188165483_0003
19/09/11 01:18:32 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_156
8188165483_0003/
19/09/11 01:18:32 INFO mapreduce.Job: Running job: job_1568188165483_0003
19/09/11 01:18:42 INFO mapreduce.Job: Job job_1568188165483_0003 running in uber mode : false
19/09/11 01:18:42 INFO mapreduce.Job:  map 0% reduce 0%
19/09/11 01:18:51 INFO mapreduce.Job:  map 100% reduce 0%
19/09/11 01:19:00 INFO mapreduce.Job:  map 100% reduce 100%
19/09/11 01:19:00 INFO mapreduce.Job: Job job_1568188165483_0003 completed successfully
19/09/11 01:19:01 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=296
                FILE: Number of bytes written=221421
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=5458319
                HDFS: Number of bytes written=197
                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)=7160
                Total time spent by all reduces in occupied slots (ms)=5893
                Total time spent by all map tasks (ms)=7160
                Total time spent by all reduce tasks (ms)=5893
                Total vcore-seconds taken by all map tasks=7160
                Total vcore-seconds taken by all reduce tasks=5893
                Total megabyte-seconds taken by all map tasks=7331840
                Total megabyte-seconds taken by all reduce tasks=6034432
        Map-Reduce Framework
                Map input records=124456
                Map output records=901325
```

```
        Map-Reduce Framework
                Map input records=124456
                Map output records=901325
                Map output bytes=7210600
                Map output materialized bytes=296
                Input split bytes=120
                Combine input records=901325
                Combine output records=29
                Reduce input groups=29
                Reduce shuffle bytes=296
                Reduce input records=29
                Reduce output records=29
                Spilled Records=58
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=166
                CPU time spent (ms)=2590
                Physical memory (bytes) snapshot=351318016
                Virtual memory (bytes) snapshot=3008946176
                Total committed heap usage (bytes)=226365440
        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=5458199
        File Output Format Counters
                Bytes Written=197
The median is: 4
```

4) C, D
5) All correct except of the C
6) A, C
7) C