PROJECT 1 REPORT

a) Transformation of data during the computations

“WordCount” is a very simple program that counts the number of occurrences of each word in an input file.

The map method in the program that takes (key, value) as parameters and processes the data in the file one line at a time, it then splits the line into tokens separated by whitespaces using the *StringTokenizer* and generates key value pairs.

The Mapper takes the raw input file and writes <”key”, 1> tuples in a while loop. Mapper finds the occurrences of the words in the input file and assigns tuples. Therefore, it maps each <”key”, value> input into one or more <”key”, value> outputs. For example, (input) <”class one is closed and class two is open”, 1> → (Mapped to :)<and, 1>, <”class”, 1>, <”class”, 1>, <”closed”, 1>, <”is”, 2>, <”is”, 1>, <”open”, 1>, <”one”, 1>, <”two”, 1>

As an option users can specify a combiner interface to do local aggregation of the intermediate outputs from the mapper on the keys. This process helps cut down the amount of data transferred from the mapper to the reducer and thus speeding up the process. The combiner deals with the output tuples from the mapper class before they make it to the reducer. The combiner pre-summarizes the tuples, therefore optimizing disk I/O in the cluster. So in our example the 100 <”the”, 1> tuples are aggregated and become one tuple <”the”, 100>. The combiner extends the reducer class and it uses *“context.write(key, result)”.* It takes tuples and tries to summarize them and writes onto the context.

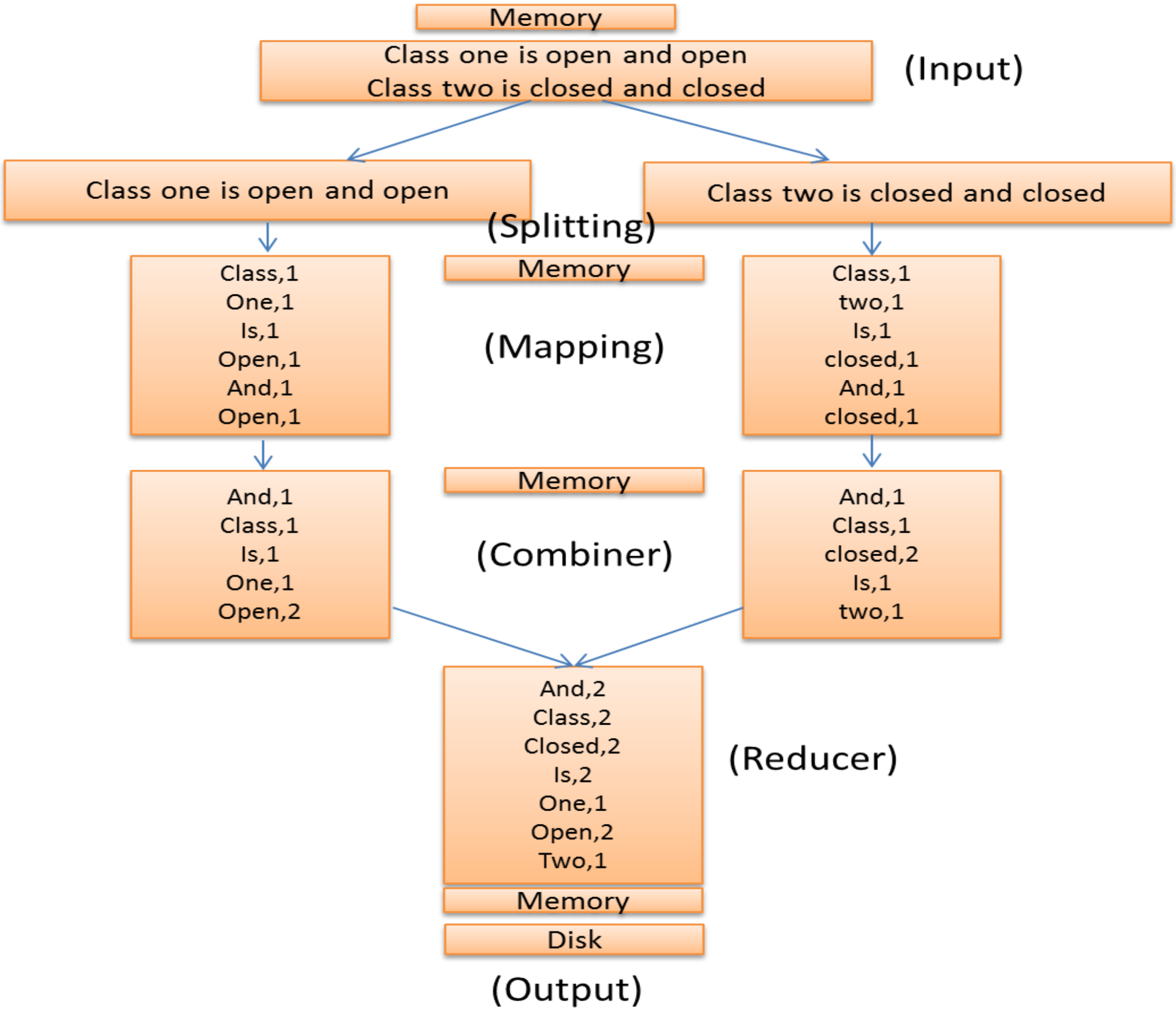
The reduce method gets a key and a value and produces the output of type *text* and *Intwritable*

Input and Output process of a Map/Reduce Job:

(input) <k1, v1> -> **map** -> <k2, v2> -> **combine** -> <k2, v2> -> **reduce** -> <k3, v3> (output)

[Hadoop documentation retrieved from http://hadoop.apache.org/docs/r0.18.3/ /mapred_tutorial.html]

b) **Figure 1:** MapReduce Word Count Process



References:

Hadoop Documentation. Hadoop Map/Reduce Tutorial, Last Published: 12/02/2008. Retrieved from: http://hadoop.apache.org/docs/r0.18.3/mapred_tutorial.html