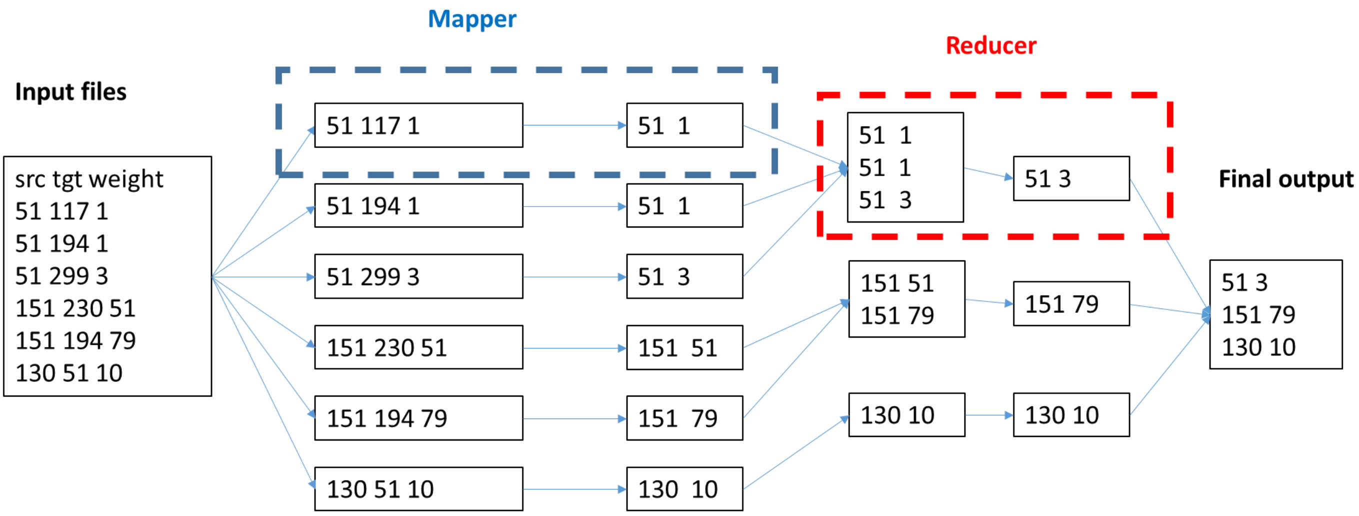
Task 1 – description

Question a



The first phase is the map phase : each line is split into 3 strings separated by a tabulation. We will keep only the source which is the first value and the weight which is the last value.

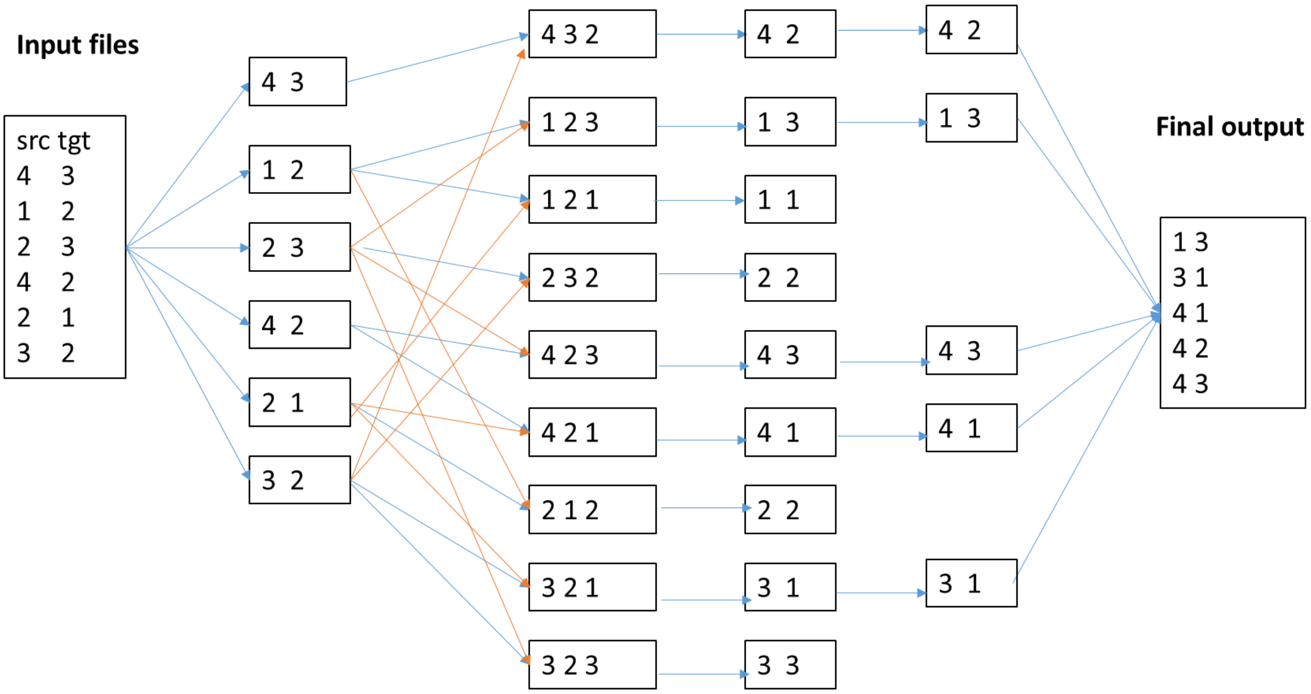
The second phase is the shuffle phae. All the data with the same source will be merged together.

Durig the reduce phase, we will keep only the highest weight for a given node.

Finally we will obtain our final output by merging the results of each reducer as illustrated.

Question b

The mapper will read each line composed of the source and the target. Then each line will be associated to the line having a source with the same value as the initial source. They can be joined on the target of the first dataset which is equal to the source of the second dataset. Then they will be reduced to the first data and the last data which correspond to the initial source and the final target. FInally the pair (v,v) are suppressed. And the final output is obtain by merging the results.


Input files
src tgt
4 3
1 2
2 3
4 2
2 1
3 2
4 3
1 2
2 3
4 2
2 1
3 2
4 3 2
1 2 3
1 2 1
2 3 2
4 2 3
4 2 1
2 1 2
3 2 1
3 2 3
4 2
1 3
1 1
2 2
4 3
4 1
2 2
3 1
3 3
4 2
1 3
4 3
4 1
3 1
Final output
1 3
3 1
4 1
4 2
4 3