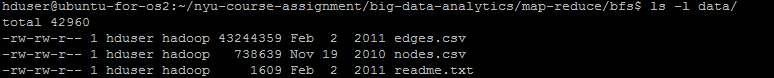
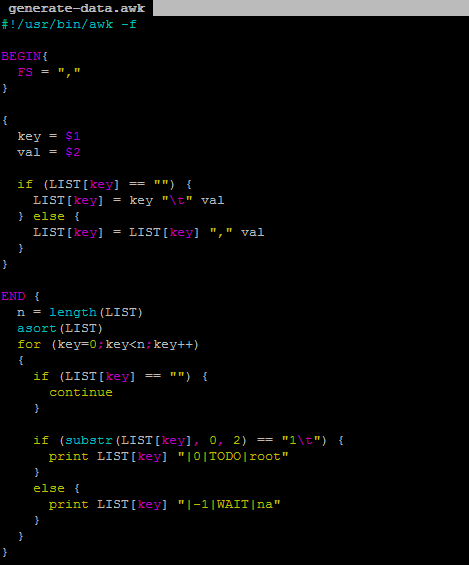
Raw data from http://socialcomputing.asu.edu/uploads/1296674602/Foursquare-dataset.zip



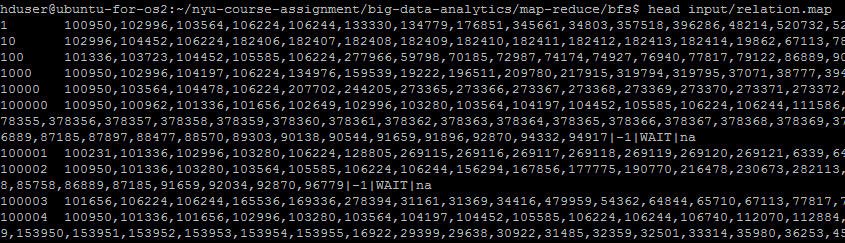
Raw data to initial input data conversion script by Awk



$ ./generate-data.awk < data/edges.csv > input/relation.map

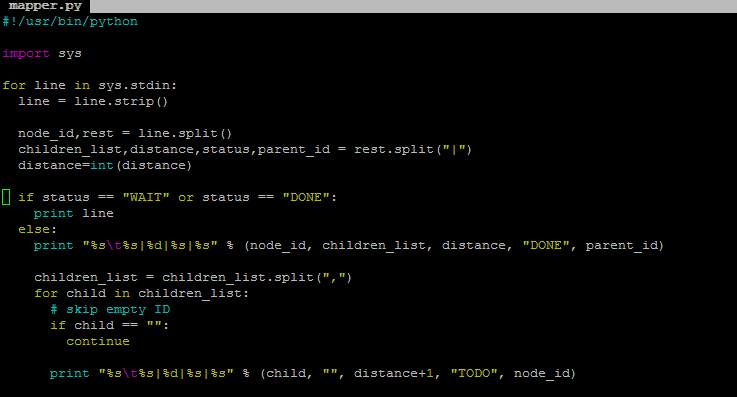


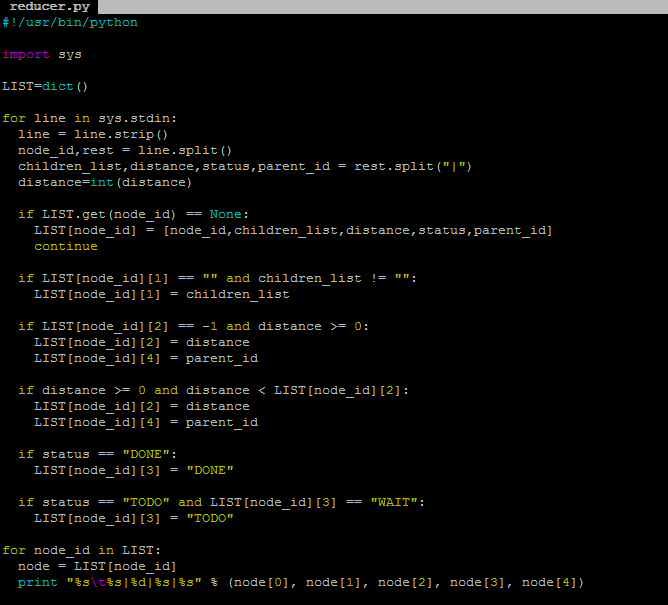
Check the input file



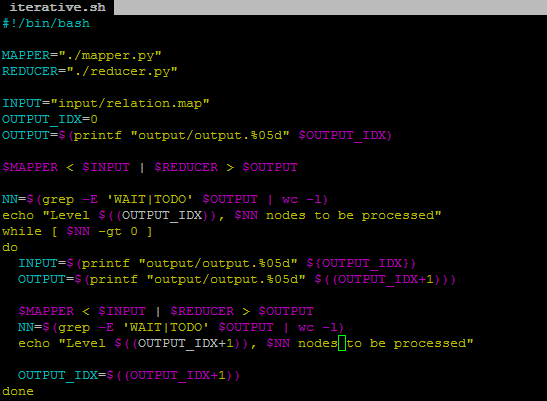
Mapper and Reducer by Python



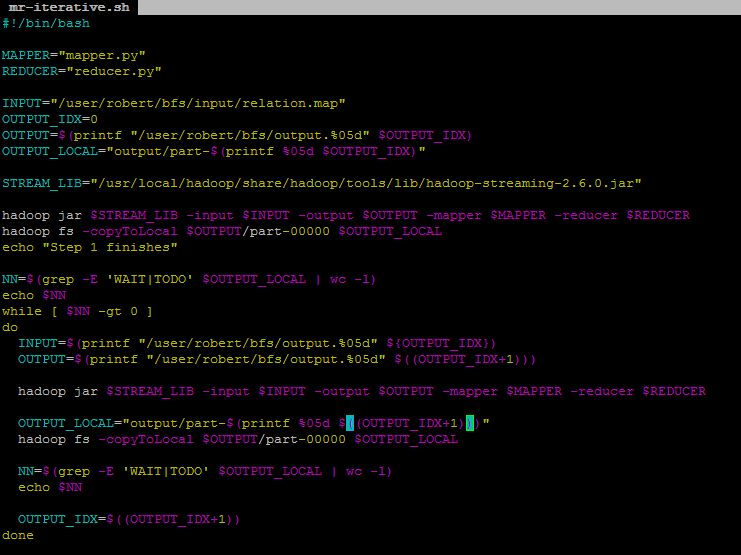




The dry-run script by bash, without HADOOP



The run script by bash as well, on HADOOP

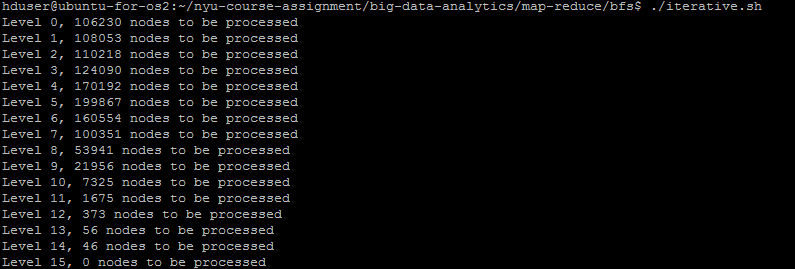


Dry run:

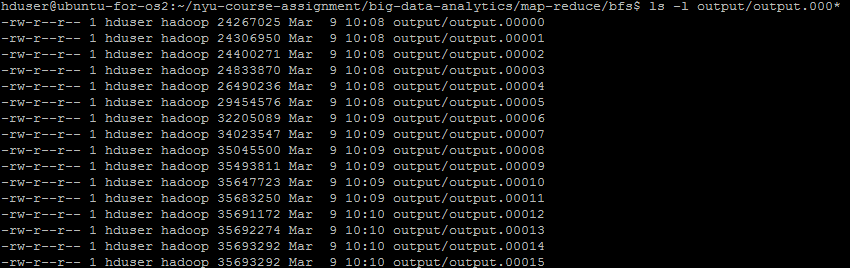
$ ./iterative.sh

Or

$ make dryrun



Dry run result:



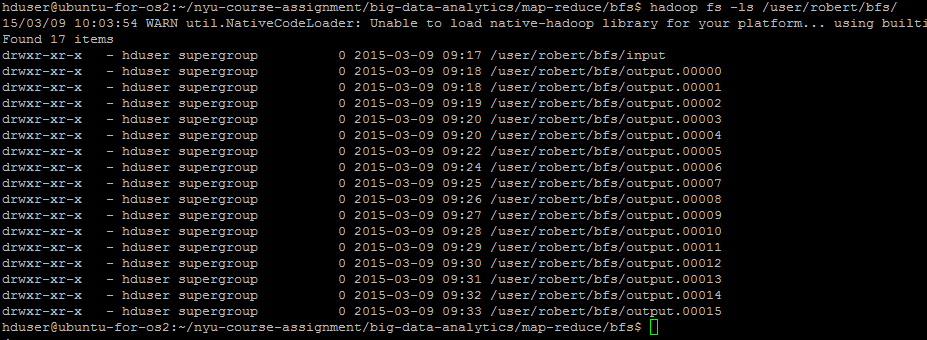
Real run:

$ ./mr-iterative.sh

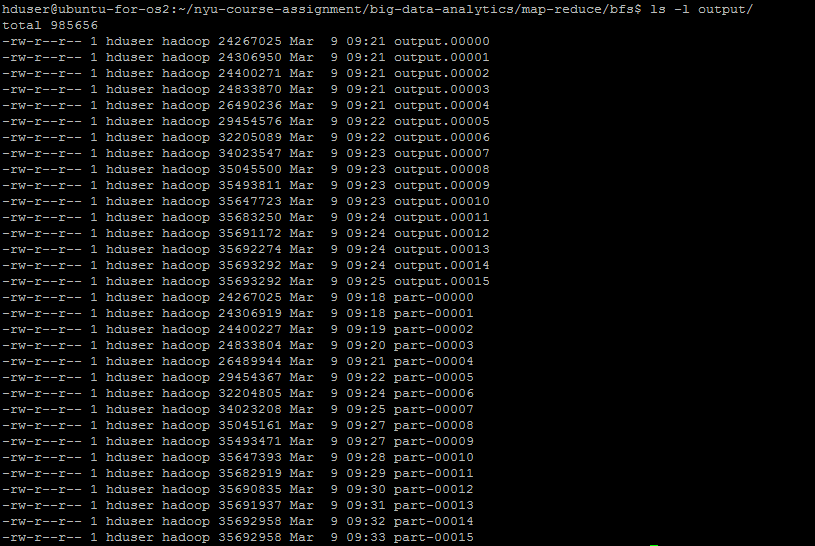
Or

$ make run

Check the result on HDFS



Check the result on local:

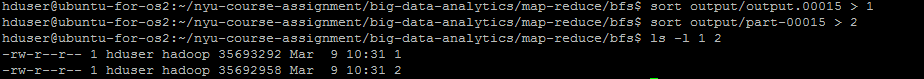


We can see both dry run results and remote results are all in the output directory, with different name convention.

The name with “part-xxxxx” are the results copied from the HDFS.

The name with “output.xxxxx” are the dry run results.

As we can see, both run have 16 iterations. We can check the part-00015 and output.00015 and see.



If we sort them and will see the size of the two files are not the same, which leaves some doubt about the correctness of the result. Further review of the two results shows the parent node for some nodes are different from the two results. One possibility is that during the shuffling, the order of the nodes are different and there are a few nodes with multiple parents but the same distance. In that case, the first node with the same distance will win.