Uni Freiburg, Web Science Group
Prof. Peter Fischer
Systems Infrastructure for Data Science - Winter 2014/15

Exercise Sheet #12: Map/Reduce, Pig and Hive

January 29, 2015

1 Map/Reduce, Pig and Hive

In order to compare Map/Reduce, Pig and Hive, we again use our social media example from the previous exercises: For each user, there is a list of users that are connected to him/her. Please perform the following steps for each of these three platforms and compare the results/effort involved

A. Perform -if necessary- schema and data modelling.

B. Compute the two-hop node neighborhood of each node, i.e. the number of (other) users that can be reached directly or by following the connections once. Each user in the neighborhood should only be counted once.

C. Create a ranking (TOP-N) of the "influential"/"popular" users, i.e. those users having the largest two-hop neighborhood

D. Roughly translate the PIG/Hive queries into a sequence or graph Map/Reduce rounds. Compare with your "native" Map/Reduce implementation. Which optimizations may be possible?

To make these tasks easier, here are some references

- Apache Pig Documentation, `http://pig.apache.org/docs/r0.11.1/`
- Programming Pig, Alan Gates, OReilly Media.
  (Online version available at: `http://chimera.labs.oreilly.com/books/1234000001811`)
- Apache Hive Website, `https://hive.apache.org/`
- Programming Hive, Edward Capriolo et al., OReilly Media.
  (Online version available at: `http://it-ebooks.info/book/941/`)