Uni Freiburg, Web Science Group
Prof. Peter Fischer
Systems Infrastructure for Data Science - Winter 2012/13

Exercise Sheet #11: Map/Reduce Joins, Pig Latin

January 25, 2013

1 Map/Reduce Joins

Map/Reduce does provide a direct implementation of the join operator. Instead, multiple approaches have been proposed to express joins in an efficient manner.

A. The most obvious choice would be a *reduce-side* join, where the map stage just prepares data and reduce performs the actual joins. Describe map and reduce funtions (in your own words or using code) and describe how the join in computed.

B. Would it also be possible to perform a job purely on the map side? What conditions would have to be fulfilled, and how would the map and reduce funtions look like?

C. What would be the respective benefits and disadvantages of *reduce-side* and *map-side* joins?

2 Pig

Express the problem setting of Sheet 10, exercise C using Pig Latin instead of pure Map-Reduce. For your convenience, the text of the exercise is included again:

Facebook has a feature which lists common friends with a person when you visit his or her profile page. In the context of this exercise let's assume that common friends will be listed for pair of persons who are friends. We would like to implement this feature using MapReduce. The friendship is a bi-directional relationship, if A is a friend of B then B is a friend of A too. Assume that you are given a file which consists of millions of lines in the following format:

PersonA [PersonB, PersonC, PersonD, ...]
...

where each line lists a person's name followed by list of his/her friends.