CS W186 Introduction to Database Systems
Fall 2019 Josh Hug

DIS 8

1 Parallel Query Processing

1. What is the difference between inter- and intra- query parallelism?

2. What are the advantages and disadvantages of organizing data by keys?

3. Given m machines with B page buffer each, along with N pages of data that doesn't have duplicates.

   (a) What is the number of passes needed to sort the data? Find the best case, in terms of N, B, and m.

   (b) What is the number of passes needed sto hash the data (once)? Find the best case, assuming that somehow the data will be uniformly distributed under the given hash function.

   (c) If you don't have a hash function that can uniformly partition the data, would round-robin partitioning be useful here? Why or why not?

(d) Instead of N pages, you are given R and S pages of data (one for each relation). What is the number of passes in order to perform sort merge join? Consider reading over either relation to be a pass.

(e) Can you use pipeline parallelism to implement this join?