Theodore Bieber

Distributed Systems

Project 3

I have pre-compiled all of the code into jar files, the instructions on how to run them are in the readme.txt file.

The part1 folder contains the source code for part1, as well as the class files and a jar, and lastly it contains the customers.txt and transactions.txt files that I used in my program.

The part3 folder contains subfolders for each query (5 total including 2a and 2b). Each subfolder contains the source code, class files, and jar files for each query.

The readme.txt file details important commands that I used to run my programs.

The output folder contains ALL of the output from the queries.

Query 2: comparing Combiner vs no Combiner

Combiner (Query2b):

Total execution time: 730 ms

Execution times for Map Task: 330 ms

Number of records sent to reducer: 50,000

No Combiner (Query2a):

Total execution time: 730 ms

Execution times for Map Tasks: 310 ms

Number of records sent to reducer: 50,000

Conclusions:

I am not seeing as large of an effect that I had initially expected, but I do see that with a combiner, less of the time is spent in the reducer. I would expect these differences to be much larger when introducing more machines into the cluster. The idea of the combiner is to reduce the amount of information being sent by condensing the data, so with more machines that need to send data between each other, we would see the differences increase.