is607WK_14_quiz

Prashant B. Bhuyan

December 16, 2014

Problem:

Your job here is NOT to load the data up. Your job is to estimate how long it would take to copy all of the data (2007 to 2014) down to a local machine, unzip it (if necessary), load the data into a fast database, process the data (e.g. perform some kind of basic summarization), and send the processed data results back to the web server. It is important that you make reasoned (and order-of-magnitude) reasonable estimates, that you explicitly state each of your assumptions, and that you present your findings in a clear format. It is not important that you come up with a correct answer. By going through this thought exercise, you'll gain some insight into the big data paradigm shift of "moving code to data instead of data to code."

Solution:

Firstly, let's assume that the internet speed is 10 Gbps and the processing is basic (sorting and grouping).

I would do the downloading and processing in parallel. Each file is approx less than 100mb so it would be much faster to sort a file than the entire data set which would potentially take weeks assuming the unzipped data is tens if not hundreds of gbs in size. As such, I would download one file and notate the time it takes to download. That file size divided by the time it took to download would give me the Download Time Per MB. Then I would simply multiply the Download Time Per MB by the Total File Size and that would give me a Download Time Estimate for how long it would take to download the entire file. Then I would repeate that process to find the Unzip Time Estimate and the Upload Time Estimate.