Charley Ferrari

IS 607 Week 14 Quiz

First, count the number of files

Dec 2007 – Nov 2014 – 7 years - 365 days – 24 hours = 61,320 total files

Average size (average of November) 64.6MB (this could easily be calculated in a more robust way, taking into account how usage grew from 2007 to 2014, but I’ll use this for the estimate.)

The example summarization I will perform is to get the total pagecounts for each language. This sort of summation is something that can be performed with each individual file, entered into a local summary databank, and then further aggregated. Doing this is an example of “moving the code to the data.” Instead of trying to get all the data in one place to perform the operation, I’m going to keep the data in chunks and perform the summary step by step. In order to further calculate the time needed, I’d work out the following:

Average time per file to download (516.8Mb / 8.43 Mbps = 61.2s)\* number of files (total: 3752784s) +

Average time per file to unzip (3s)\* number of files (total: 183960s) +

Average time to load each file into fast database (18ns)\* number of files (total: 1.1ms) +

Average time to process the summation for the hour in question (12ms), and enter summation into another database (18ns, using same load time as above). \* number of files (735.8s)

Total time to sum the aggregate database (12ms, using same database calculation time as above). This is done in one database, not across multiple files.

Total time to send the result to a web server (once again, done in the one summary database.)

The last two totals are not multiplied by the number of total files, and are relatively small operations, so I can just add up what is done for each file. Adding up the items above, we get 3752784s + 183960s + 735.8s = 3937479.8s (I skipped over the 1.1ms since our other sums were orders of magnitude larger.)

This is a very rough estimation, but does correctly show that it’s a better idea to break the operation into chunks, and then get a total calculation from the summaries of the components.