Instead of sending you a bunch of different messages on Slack I thought it would be better to put it all in this word document.

Finished my experiment on the 10gb TPCH-Database. I would like you to look at a couple certain things and see if I did the formulas correct. First thing to look at is the Speed Up Column. I used the formula as follows.

(AVG Time SOURCE /AVG Time SAMPLE) – 1 So for the first query it was

14.35/8.21 = 1.7478

1.7478 – 1 = .7478

.7478 \* 100 = 74.78%

Which Excel had 74.79 so close enough.

Speed up = AVG Time SOURCE /AVG Time SAMPLE

no need to subtract 1

This is the formula I used for all my Speed Up Results. Now if you follow down to the bottom you will see final results. Under Average Speed Up I have 158.73%. Now I just used the formula I used for all the other speed up results, again, shown above on the final results column. If I attempt to average out the speed ups by just adding up the speed up columns and dividing by 15 I get another number 210%. Which number is better to go with? I felt the 158% because it’s a true Speed Up Percentage of all the Result Averages.

Next column to look at is the Average Relative Error. I am not worried about any mathematical calculation errors or anything, and if you look at the results it looks like it worked fine and had some great numbers. The problem I came out with is how to show it in chart form. Someone who does not know relative error might think low is bad or that those numbers are high. I put it in chart form (sheet 2 on excel sheet) but with the numbers entered in the chart will only go to 10 (which makes the numbers look bad) and I could not figure out how to change the dataset to 100 (double click on the Y axis numbers and in right hand side “format axis” panel in excel change 10 to 100. You can google for these questions.).

Relative error = | source\_answer – sample\_estimator | / source\_answer \* 100%

The lower the better for relative error

Finally I believe from everything that I read and what we have discussed so far my thesis is going to show that CS2 worked on normal data but was never tested on big data. So, do you think I should create a whole second results section from maybe a 1GB database on hive? Just wipe the original database and throw on a 1GB tpch database in which I can create the same samples and same queries to support my results on the 10GB database?

Your current result is on 1GB as the database is called tpch1g. How about if you have finished the thesis writing and if you still have time, I will create you a 10GB dataset in hive and you can do a test if you are interested?

The reason I ask this is I read a little of Tyler’s thesis and noticed he is doing a comparison of sorts in it. And your CS2 paper also is comparing against TS and TUG. But I don’t really know what I am comparing to, to give me a more solid thesis showing the results. I am showing that yes, CS2 does work on big data, and going to show the experiment. But I am not really doing any comparisons (besides showing how the sample data compares to the source data).

And on a final note Sheet 2 I was able to create 3 charts that I could think of based of the information I had. The first one shows the difference between the source and sample tables in time it took to finish the queries. The second chart shows the Relative Error from all the tests (this is the one I would like to change the data range to 100 instead of 10) and the final one is the speed up table showing the increase in overall speed from the source table to the sample table. If you think any of this will not work or want me to make tweaks on anything please let me know and I will get right to it. Just looking at Tyler’s Thesis I feel as though I have nowhere near the amount of testing that he had (I know its two totally different subjects and completely different data, but he has 21 pages of information on his tests, where I feel like I will only have 10)

I just want to start writing the thesis on a good foot. I know I am showing CS2 works on big data, and to use my tests for my results section. What about the body of the paper before the tests. I know I should discuss CS2 with normal data and JS and TUG, but after that what should I discuss? Big data? Query optimization? I just see like 35+ pages ahead of me to write and I am not sure where I should get the information to help fill it (besides the CS2 and tests)

You should come up with a draft outline first. Please look at Tyler’s paper for a structural example. Here is a suggestion:

* Abstract
* Introduction
  + Talk about query optimization, query size estimation
  + Histogram, sample, and synopses (JS, TuG)
* Background
  + Join graph of a database
  + Join graph
  + Big data
    - Differences of Hadoop distributed file system (HDFS) vs centralized files system
* CS2
  + Briefly talk about what is CS2
  + Construction of CS2
    - Sampling on source relation
    - Correlated sampling
* CS2 on big data
  + We are implementing CS2 on a big data file system (HDFS)
  + Estimation of big join queries using CS2 (same formula as CS2 on centralized file system)
* Experiments
  + Experiment setup (machine, dataset, operating systems, testing queries)
  + Results
    - Tables
    - Figures
    - Result analysis: CS2 is faster
* Conclusion
* References