Writing Assignment
Spring 2014
CMSC 362
Marmorstein
First Draft Due: Apr. 4
Final Version Due: Apr. 18

Background and Introduction

One of the most important subfields of database theory is something called “query optimization”. When you are working with a large dataset, unoptimized queries can take a long time. For many applications, this delay is unacceptable. One solution to this is to buy faster hardware or switch to a different database architecture (such as a NoSQL solution) that allows parallelization. Often, though, a slow query can be made faster simply by rewriting it, by adding an index, or by modifying the schema of the database. Sometimes this can be done automatically by the database engine (which is the focus of most research in this area), but for this paper, I want you to do some analysis “by hand”.

Stack Overflow is a popular forum for discussion of technical topics. Since 2009, they have published a downloadable archive of all their content (suitably anonymized to protect privacy) which you can obtain at this link:

http://blog.stackoverflow.com/category/cc-wiki-dump/

For this paper, I want you to:

1. Download the stack overflow archive.

2. Come up with a set of queries (at least 3) against that data that take more than a few seconds to execute.

3. Try to find ways to speed up those queries.

4. Write up your results in a 3-5 page research paper.

Format of the paper

Please make sure your paper is double spaced in two-column form with 0.5 inch margins and in a readable font (it should be a size between 10pt and 14pt). It should be printed in portrait orientation on 8.5x11” paper. The gutter between the columns should be no more than 0.25 inches. You should also include a title page which contains only the title, your name, and the words “CMSC 362: Spring 2014” at the front of your paper. The title page does not count toward the page limit.

Evaluation

You will receive a grade on this assignment based on the following criteria:

(10 pts) Organization

Your paper must be properly structured as a computer science research paper. **It should contain numbered or titled sections** for at least the following: an introduction, a previous work section (i.e. a literature review), a methodology or implementation section, a results section, and a conclusion. The introduction should outline the motivation for your project (the problem you were assigned to solve), introduce any technical background the reader needs to understand your work, and give a roadmap of the rest of the paper (if you don't know what a roadmap is, feel free to visit me during office hours). The previous work section should cite relevant papers in which others have tried to solve the problem and explain why more work is needed in the field (yes, this means

you will need to go to the library and find some papers on database optimization). The methodology section should give a detailed explanation of how you solved the problem. In the results section, you should analyze how well your solution worked. You should use appropriate graphics to depict your data and summarize/explain those graphics in the text. The conclusion of the paper should address areas in which your work could be improved or built upon for future research.

It is not enough to simply write a sentence or two for each of these parts. You will lose points if these sections are not thorough and complete. Pay particular attention to the previous work section, which MUST contain references to other work on database performance. Be sure to cite your references carefully. Failing to give credit to others is plagiarism and is an honor code violation.

(15 pts) Data Collection

The bulk of your grade will come from providing accurate data in a form which is easy to interpret. You should use a sound and rigorous research methodology. You must provide a comprehensive performance evaluation by conducting carefully timed experiments (the Unix "time" command is probably your friend here). You must present these results in a way I can understand. Appropriate use of figures and charts to present your data is very important for this.

(15 pts) Content

I will also evaluate your paper based on the content of your prose and the depth of your analysis. Your explanation, both of your own work and of others, must be sufficiently detailed and factually correct. You should include enough detail that, given time and access to the data set, I could repeat your experiment in the Hardy House lab.

(10 pts) Spelling, Grammar, style and Appearance

Your paper must follow proper English diction and syntax. Avoid passive voice, run-on sentences, and other grammar errors. Your writing must avoid the use of jargon and informal language. You should not assume that your reader is a member of our class or knows you personally. You should assume that they are familiar with the basic principles and vocabulary of database theory. Be careful not to use the first person singular.

Submitting

I would like you to submit an electronic draft of the paper by **5pm** on April 4th. There will be a submit link on the course web site. The paper MUST be in .odt (OpenDocument) or .tex (LateX) format. I will not accept .doc, .docx, or .pdf files. (This is so I can easily mark it up and suggest changes).

I would like you to turn in a **hard copy** of the FINAL draft paper by **5pm** on April 18th. You should deliver your paper to my office. (If I am not in my office, please slip it under the door). The paper should be properly stapled so that I can easily grade it.