Database Systems Technology (CSC443-W2019)
Assignment 1

Submission Deadline: Feb 10, 2019 11:59 PM

1 Cost Analysis - 35%

Calculate the cost of operations given in the table in **Slide 19** of **"Storage and Indexing"** lecture. All calculations should be based on parameters **B, R, D** as defined in the lecture and you should make exactly the same assumptions as given there.

You need to show your work how you arrived to those results. The cost is the expected cost. For range operations you should consider the worst case that range can be as wide as the entire dataset.

2 Verification Using Real Data Files - 65%

In this section, you will create a SQLite data file and you will write a program to perform the query operations in the "Cost of Operations" table (mentioned in the first part of the assignment). You need to follow the steps below:

1. Download the sample CSV file (zipped) from this link (http://eforexcel.com/wp/wp-content/uploads/2017/07/500000-Records.zip). This is an artificially generated sample employee list with many fields. We mostly care about **Emp ID** column but you need to load all columns to the data file in the next step.

2. Download and install SQLite 3.26.0 (the latest version) on your machine. SQLite is a C-language library that implements a small, fast, self-contained, high-reliability, full-featured, SQL database engine. SQLite is the most used database engine in the world. It is built into all mobile phones and most computers and comes bundled inside countless other applications that people use every day.

3. Write a program or a script to create a SQLite database by loading the CSV file as a table named "Employee". Insert the rows with the same order as in the CSV file. For simplicity, define all the columns, except **Emp ID** as fixed-length strings - CHAR(len) (first calculate the maximum

length of each column). The **Emp ID** should be defined as INT. This will make the records fixed-length. You need to repeat this process four times to create four different databases with the following characteristics:

(a) Without any index with page size of 4KB

(b) Without any index but with page size of 16KB bytes

(c) With primary index on "Emp ID" column (Unclusterd Index) with page size of 4KB

(d) With primary index on "Emp ID" column but defined as clustered (use CREATE INDEX WITHOUT ROWID) with page size of 4KB

For more information on how to do the above tasks, please refer to SQLite documentation: https://www.sqlite.org/docs.html

4. Write a program to read the data files and make the following queries. Please note that you cannot use any of SQLite API functions. You need to understand the file format of the SQLIte data file (https://www.sqlite.org/fileformat.html) and then use file read functions (Python or C++) to read directly from the file. You shouldn't use any buffer in memory and you need to read data page by page each time. The file format may look complex but in your special case you can ignore many of them (e.g. WAL, Schema, ...) and you don't need to deal with them.

   (a) Query and print the employee id and full name of anybody whose last name is "Rowe" (this will be a Scan operation)

   (b) Query and print the full name of employee #181162 (this is an Equality search)

   (c) Query and print the employee id and full name of all employees with "Emp ID" between #171800 and #171899 (This is a Range search)

   All the queries need to be run for all four data files. Please note that for data files with index, you need to use the indexes (search in index first and then get data).

5. Your program should count and print the total number of page reads (from file) for every page type (i.e. header, data, index internal node, index leaf node) as well as the average time to read one page.

6. Compare your results with the formulas you obtained in Q1. Are they consistent? If not, what could be the reason?

3 Submission

You need to submit the following to MarkUs before deadline:

1. A report including the answers to Q1 and Q2 and all the results from Q2. You need to include the screenshots of when you run your programs and the output that is printed. The report should be in PDF format.

2. You need to attach the source code of all scripts and programs. They should be written modular and should be self-documented.

3. Do not include the generated data files.