McMaster University
CS 3DB3 Fall 2017
Assignment 2
Due: November 3, 2017 at 10:00am

October 6, 2017

For this assignment, you will continue to work with *Hamilton Street Railway (HSR)*, and perform data analytics over real data, tracking bus routes over the period May 1 - May 7, 2017. On Avenue, under Assignments, you will find:

a) the DDL for CREATE TABLE statements, `createTables.ddl` (to create the necessary tables), and

b) the INSERT statements `loadData.ddl` (to load data into the tables).

These files correspond to a simplified schema shown in the E-R diagram `asg2ER.pdf`. Please run these two scripts on your database (i.e., remember to update the CONNECT TO statement with your database name). You will use this schema for the questions below.

I. SQL (65 marks)

Write and provide SQL statements for the following queries. Execute each of your SQL queries against your HSR database, and give the result of each query.

q1) [3 marks] Identify all buses (busID, age, manufacturer) with advertising revenue greater than $9000.

q2) [3 marks] Find the number of students in the database. A student is defined as a person with a student occupation or is less than 25 years old. Do not include duplicates. (Hint: If needed, you may use the date() function.)

q3) [4 marks] How many students took bus route #5 on May 3rd, 2017?

q4) [4 marks] For each bus route, find the total advertising revenue. Return the bus route number and the total revenue. Order the results in descending order of total revenue.

q5) (a) [4 marks] Find all drivers who have less than 3 infractions. Return the driver's ID, first name and last name.

(b) [5 marks] For each driver, return the total demerit points and total fines incurred. Do not include drivers with less than 2 demerit points in the result. Sort the result such that the most offending drivers (in terms of demerit points, total fines) are listed first.

q6) [4 marks] Determine those buses that are the unique (only) bus made by their manufacturer. Return the busID and the manufacturer.

q7) (a) [4 marks] For each passenger type, find the total fare revenues. Return the passenger type and the total (fare) revenue.

(b) [2 marks] Extend your query in part (a) to only list passenger types with revenues greater than $500.

(c) [2 marks] Extend your query in part (a) to return the most profitable passenger type (in terms of total fare revenue) on May 1, 2017.

q8) (a) [4 marks] Determine the most popular bus route on May 7, 2017 (according to the number of passengers). Return the route ID and the number of passenger trips.

(b) [4 marks] Which day contained the largest volume of passenger trips? Return the date and the number of trips taken.

q9) [4 marks] Find all persons who visited a library on either May 5, 2017 or May 6, 2017. Return the person's occupation. Do not include duplicates.

q10) [5 marks] Find the drivers who have worked with HSR for more than 5 years, with a salary greater than $80000, and with less than 10 demerit points on their driving record. Return the driver's first name, last name, and ID.

q11) [6 marks] Find all students who attended the "Marauders Tennis" match and arrived via a bus on route 4. Return the student's first name, last name, and their gender.

q12) [7 marks] Assuming that the bus schedule has not changed since May 2017. Suppose you would like to attend the "Marauders Basketball" game (an event). The game starts at 5pm, and you'd like to arrive at the site between 4:20pm and 4:50pm, which routes can you take? Use May 1, 2017 as a reference date for the bus schedule information. List the route ID, the bus stop name (where you should get off the bus), and the scheduled arrival time.

II. Relational Algebra (35 marks)

For each of the SQL queries in Part I, give the corresponding relational algebra expression.

III. Indexes (20 marks)

For the workload given in Part I, several queries have been showing poor performance (i.e., increasing response times). Your task is to improve the performance of this workload as much as possible by recommending four indexes that should be defined on the tables. What four indexes would you recommend? For each index, state:

- The attribute(s) the index is defined on.

- Properties of the index (e.g., type of index, clustered/unclustered, etc.)
- Which queries (q1 - q12) you think this index will help, and why.

Grading

This assignment is worth 14% towards your final grade.

Submission

All files are to be submitted using the Avenue system. Please ensure you submit all files with the correct names, as described below. In each file, include your name and student ID number. Upload four files with the indicated file extensions (no compression based `.tar, .zip, .rar` files).

- For Part I: Submit your SQL statements and the result for each query in two files. Submit your SQL statements in a script file called `queries.sql`, and the corresponding query results in a file called `queries.results`. Ensure your SQL statements are syntactically correct and that they are executable on the DB2 servers. Non-executable queries will not be marked. Clearly label, with comments, which query the result tuples correspond to in `queries.results`.
- For Part II: Submit your relational algebra expressions in a file named `ra.pdf`.
- For Part III: Submit your index recommendations in a file named `index.pdf`.