Virginia Tech.
Computer Science

CS 4604 – Introduction to DBMS
Spring 2015, Prakash

Homework 1: Relational Algebra and SQL

(due February 9th, 2014, 4:00pm, in class—hard-copy please)

Reminders:

a. Out of 100 points. Contains 5 pages.
b. Rough time-estimates: 3~6 hours.
c. Each HW is supposed to be done individually.
d. Please type your answers. Illegible handwriting may get no points, at the discretion of the grader. Only drawings may be hand-drawn, as long as they are neat and legible.
e. There could be more than one correct answer. We shall accept them all.
f. Whenever you are making an assumption, please state it clearly.
g. Unless otherwise mentioned, you may use any SQL/RA operator seen in class/in textbook.
h. Unless otherwise specified, assume set-semantics for RA and bag-semantics for SQL.
i. Feel free to use the linear notation for RA and create intermediate views for SQL.
j. Lead TA for this HW: Elaheh Raisi.

Q1. RA: Warming up [18 points]

Consider the following two tables, T1 and T2:

T1		
P	Q	R
10	a	5
15	b	8
25	a	6

T2		
A	B	C
10	b	6
25	c	3
10	b	5

Show the results of the following relational algebra queries (3 points each):

Q1.1. $\pi_Q\,(T1) \cap \pi_B\,(T2)$

Q1.2. $\sigma_{R>5 \vee Q=a}(T1)$

Q1.3. T1 $\bowtie_{T1.P=T2.A}$ T2

Q1.4. T1 $\bowtie_{T1.Q=T2.B}$ T2

Q1.5. T1 $\bowtie$ T2 (assume the natural join happens for columns P and A)

Q1.6. T1 $\bowtie_{T1.P=T2.A \wedge T1.R=T2.C}$ T2

Q2. RA: Minimal set [6 points]

The core operators in relational algebra are selection (σ), projection (π), cross product ($\times$), union ($\cup$), and difference ($-$). Show that the selection operator is necessary; that is, some queries that use the selection operator cannot be expressed using any combination of the other core operators.

Hint: It is sufficient to argue with a specific example. Consider a relation R with the schema R = (A: integer), and a particular instance of R with only two records [0, 1]. Note that $\sigma_{A=0}(R)$ gives a relation Rs = (A: integer) and its instance is [0]. Can you argue that we cannot get the same result with any combination of the other four operators?

Q3. RA: Bars [25 points]

Consider the following relational database that stores information about bars and customers (keys are underlined, field types are omitted):

Drinker (name, address)
Bar (name, address)
Beer (name, brewer)
Frequents (drinker, bar, times a week)
Likes (drinker, beer)
Serves (bar, beer, price)

Write the following queries in relational algebra:

Q3.1. (4 points) Find all drinkers who frequent James Joyce Pub.

Q3.2. (4 points) Find all bars that serve both Amstel and Corona.

Q3.3. (5 points) Find all bars that serve at least one of the beers Amy likes for no more than \$2.50.

Q3.4. (6 points) For each bar, find all beers served at this bar that are liked by none of the drinkers who frequent that bar.

Q3.5. (6 points) Find all drinkers who frequent *every* bar that serves some beers they like.

Q4. SQL: The School DB [21 points]

The schema of the database is provided below (keys are underlined, field types are omitted):

student(sid, sname, sex, age, year, gpa)

dept(dname, numphds)
prof(pname, dname)
course(cno, cname, dname)
major(dname, sid)
section(dname, cno, sectno, pname)
enroll(sid, grade, dname, cno, sectno)

In this assignment, you will only deal with querying part of SQL. You are NOT allowed to tamper with (change the contents of) the database, i.e., CREATE, INSERT, DELETE, ALTER, UPDATE etc.

Write SQL queries that answer the questions below (one query per question). The query answers must not contain duplicates, but you should use the SQL keyword distinct only when necessary. For this question, creation of temporary tables is NOT allowed, i.e., for each question you have to write exactly one SQL statement (possible using nested SQL).

Q4.1. (2 points) Find the names and gpas of the students who are enrolled in 312.

Q4.2. (2 points) Find the name of the oldest student.

Q4.3. (3 points) Find the names and majors of students who are taking one of the Artificial Intelligence courses.

Q4.4. (3 points) Find the names of students who are enrolled in a course from both the "Computer Sciences" and "Chemical Engineering" departments.

Q4.5. (5 points) How many students have more than one major? (*Hint*: requires a nested query)

Q4.6. (5 points) Find the name(s) of the oldest first year student {year = 1} (*Hint*: requires a nested query)

Q5. SQLite: Storing Employee Information [30 points]

This question is on an employee salaries database that stores information about employees in Montgomery County, MD. Download and install SQLite3 from http://www.sqlite.org. Feel free to use SQLite3 for testing and practicing SQL queries in general.

Warm-up
Follow the documentation and load the small sample database at:
http://courses.cs.vt.edu/~cs4604/Spring15/homeworks/hw1/cs4604-hw1.db

It has a table “Employees” which includes Full Name, Current annual salary, gross pay received, overtime pay, Department (This is the abbreviation of department name; for example BOA stands for Board of Appeals,…), Division, Assignment Category (Parttime-Regular, Fulltime-Regular), Position Title (Supervisor, Manager, Social worker, Library assistant,..) and Date First Hired.

As a sanity check that you have the correct database, running the following command at a Unix/Linux/Cygwin prompt:

```
your-machine% sqlite3 cs4604-hw1.db ’select count(*) from Employees;’
```

should return

```
(40)
```

We want to write SQL queries to do the following:

- Query1: Return part-time employees
  Hint: you have to look at the Assignment Category column.

- Query2: Return the total number of departments in the database.

Larger CSV file
A bigger raw comma separated value (csv) file is given here:
http://courses.cs.vt.edu/~cs4604/Spring15/homeworks/hw1/employee-sample.csv

It is a subset from the 2013 employee salaries dataset (if you are curious, the official dataset is at http://catalog.data.gov/dataset/employee-salaries-2013). We want to write queries to do the following:

- Query3: List people who work in the same department as “Adcock Sr Gerald W”.
- Query4: List all the departments and total number of employees working in that department.

Life without SQL
Finally, in your favorite language (Python/Perl/Ruby/Java/C++ etc.) write code to do both queries above (Query3 and Query4) on the csv data file directly. Notice: the end-of-line convention is the DOS one (CRLF).

Deliverables

Q5.1. (2 points) The SQL query for Query1.

Q5.2. (2 points) The SQL query for Query2.

Q5.3. (2 points) The output of running Query1 in SQLite on the sample database.

Q5.4. (2 points) The output of running Query2 in SQLite on the sample database.

Q5.5. (4 points) The SQL query for Query3.

Q5.6. (4 points) The SQL query for Query4.

Q5.7. (3 points) The output of running Query3 on the csv file after loading it in SQLite.

Q5.8. (3 points) The output of running Query4 on the csv file after loading it in SQLite.

Q5.9. (4 points) Hard copy of your python/perl/etc code for doing Query3 on the raw csv file directly.

Q5.10. (4 points) Hard copy of your python/perl/etc code for doing Query4 on the raw csv file directly.

Hints

For loading the csv file,

- Again, the end-of-line convention follows the DOS format (CR LF).
- Use the .import and .mode csv commands of sqlite3 or check the link: http://hackgeo.com/foss/sqlite-how-to-import-csv
- A cheat sheet for sqlite3 commands http://www.natontesting.com/2008/02/09/sqlite3-cheat-sheet/
- Again as a sanity check, the command

  ```
  your-machine% wc -l  dblp-sample.csv
  ```

  should return

  ```
  1871 dblp-sample.csv
  ```