CARNEGIE MELLON UNIVERSITY
DEPARTMENT OF COMPUTER SCIENCE
15-415/615 - DATABASE APPLICATIONS
C. FALOUTSOS & A. PAVLO, FALL 2015

Homework 1 (by Dana Van Aken)
Due: <u>**hard copy, in class**</u> at <u>**3:00pm, on Monday, Sep. 21**</u>

VERY IMPORTANT: Deposit **hard copy** of your answers, in class. For ease of grading, please

1. **Separate** your answers, on different page(s) for each question (staple additional pages, if needed).
2. **Type** the full info on **each** page: your **name**, **Andrew ID**, **course#**, **Homework#**, **Question#** on each of the 5 pages.

Reminders:

- *Plagiarism*: Homework is to be completed *individually*.
- *Typeset* all of your answers whenever possible. Illegible handwriting may get zero points, at the discretion of the graders.
- *Late homeworks*: in that case, please email it
  - to all TAs
  - with the subject line exactly `15-415 Homework Submission (HW 1)`
  - and the count of slip-days you are using.

For your information:

- Graded out of **100** points; **5** questions total
- Rough time estimate: *approx. 6 hours* - 1 to 2 hours per question

Question	Points	Score
Entity-Relationship Diagram	25	
SQL Tables from the ER Model	15	
Relational Algebra for a Q & A Website	30	
Relational Tuple Calculus (RTC)	15	
Relational Domain Calculus (RDC)	15	
Total:	100	

Question 1: Entity-Relationship Diagram . . . . [25 points]

On separate page, with '[course-id] [hw#] [question#] [andrew-id] [your-name]'

Consider a database to store information about a Sports Organization. The database has the following properties:

- Each league has a name and unique league ID (integer). The 'NFL' (National Football League) is an example of a sports league.
- A league may be affiliated with one or more teams.
- Every team is affiliated with exactly one league.
- We store the name, colors, and unique team ID (integer) for each team. For example, the name of Pittsburgh's football team is the 'Steelers' and their team colors are 'black & gold'.
- A team has exactly one home stadium. A stadium may be a home to zero or more teams. For example, the home stadium of the Pittsburgh's baseball team, the Pirates, is 'PNC Park'.
- For home stadiums, we record a city, name, and unique stadium ID (integer).
- A team has at least one player, and a player plays for exactly one team.
- A team has exactly one manager, and a manager manages exactly one team.
- Every player has a number and a position. For example, Pittsburgh Steelers player Ben Roethlisberger has the number 7 and his position is 'quarterback'.
- For managers, we record the dollar amount of their annual bonus.
- Both players and managers are types of employees.
- For each employee, we store a name, salary, and unique employee ID (integer).

Given this description of the database and its constraints, we have created a mostly correct Entity-Relationship Diagram, shown in Figure 1. This diagram is at this link - feel free to use it as a starting point.
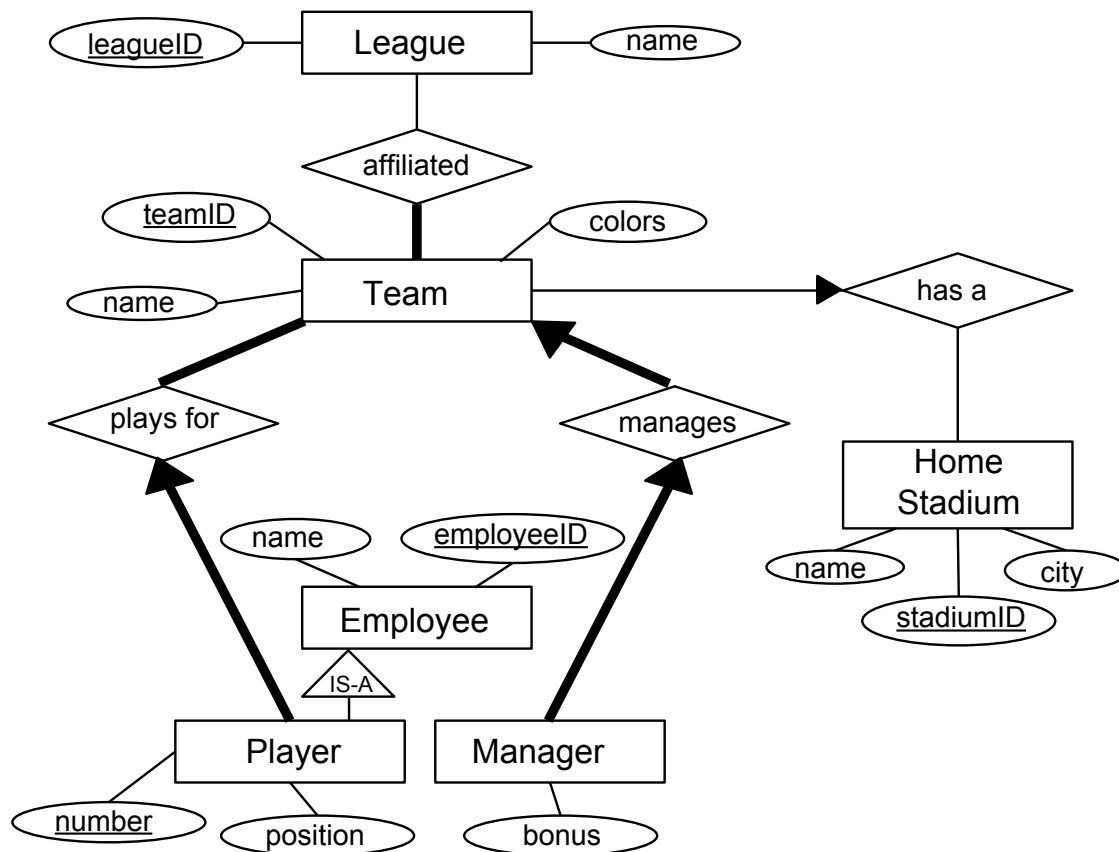
Figure 1: Almost correct ER diagram

(a) [**10 points**] Find and correct any mistakes in the given ER diagram. Specifically, number and list them, like, e.g.

  1. delete: arrow, from x to y
  2. change to bold line: thin line, from z to w
  3. change to bold box: entity e

(b) [**5 points**] There may also be some missing element(s). If none, say 'none' - otherwise, add them to the picture, **and** list them, numbered. E.g.

  1. add: attribute a, to entity e
  2. add: bold line, arrow, from c to d.
  3. add: weak entity, f, with attributes ….

(c) [**10 points**] List and number all the bold lines and all the arrows that are in the final, corrected version of the diagram. E.g.

  1. **bold**, line, from `Department` to `Employee`
  2. thin, **arrow**, from x to y

Clarifications/Hints:

• List your assumptions, if any. We will accept all reasonable assumptions.

Question 2: SQL Tables from the ER Model . . . . [15 points]

On separate page, with '[course-id] [hw#] [question#] [andrew-id] [your-name]'

Consider a database for a Twitter-like organization. It records information about users, followers, and Tweets. The constraints are exactly as shown in Figure 2. Users and Tweets have unique identifiers as shown in the Figure, with binary relationships among them as illustrated. To clarify:

- The arrow from "Tweet" to "User", is thick.
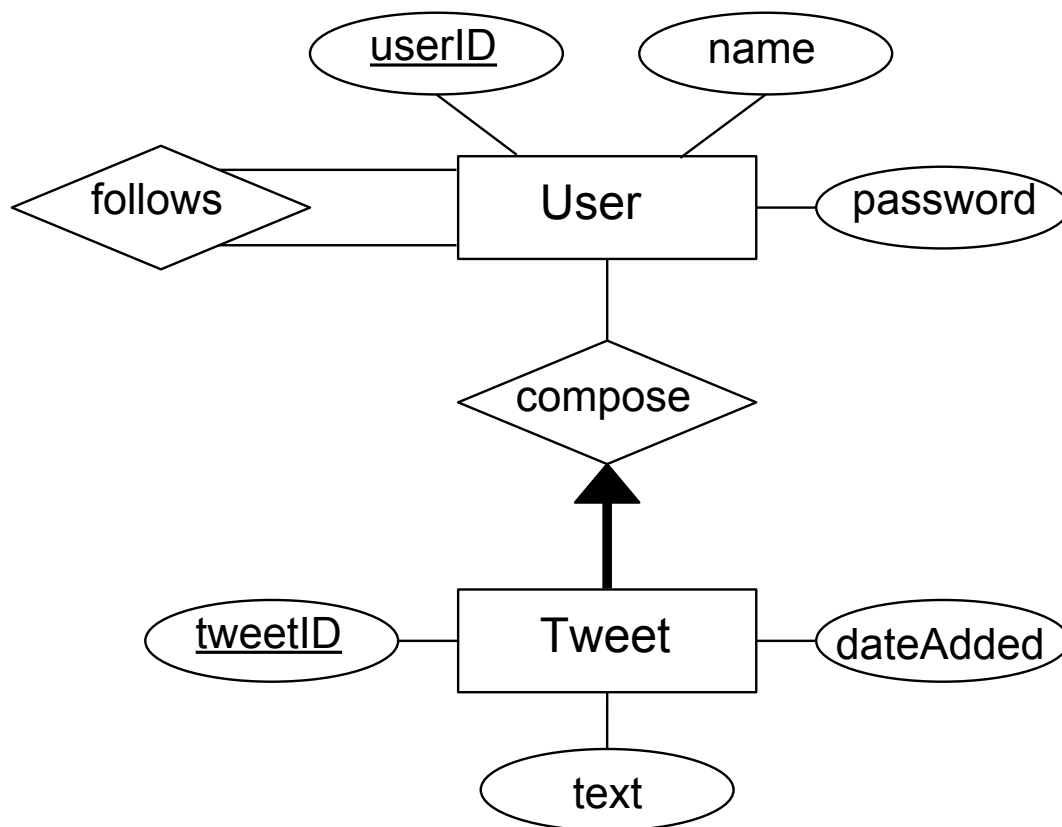- No other lines, boxes, or diamonds, are thick.

Figure 2: ER diagram for Twitter: turn to SQL tables

(a) [**15 points**] Give the DDL statements, that correspond to the above ER diagram.
- Use proper data types (we'll accept all reasonable choices).
- Avoid syntax errors (we'll forgive missing semicolons).
- Specify your decisions with respect to CASCADE deletions.
- **Without** using CHECK statements, enforce as many as possible of the implied integrity constraints as you can.

Question 3: Relational Algebra for a Q & A Website. [30 points]

On separate page, with '[course-id] [hw#] [question#] [andrew-id] [your-name]'

Consider the relations of a database for a `Question & Answer Website` as shown in Table 1. These relations describe a Q & A discussion board, where users can ask questions, post answers, and vote for the most helpful answers (as in StackExchange(TM)).

userID	name	reputation
U0	Judy	44
U1	Will	251
U2	Anne	11
U3	Gary	9
U4	Erika	200
U5	Mike	75

(a) `UserProfile` Table

questionID	userID	title
Q0	U1	"Making a one-to-many relation lazy"
Q1	U4	"Django = DatabaseError: No such table"
Q2	U5	"Inefficient database query inside a for loop"
Q3	U5	"Help with 15-415 homework…"

(b) `Question` Table: who posed what question

questionID	userID	vote_count
Q0	U3	46
Q1	U0	25
Q1	U2	131
Q2	U5	5
Q2	U3	21
Q2	U4	118
Q3	U5	45
Q3	U1	87

(c) `Answer` Table: who answered what question; how many people voted for that answer

Table 1: Relations of a database for a `Q & A Website`.

We have the following tables:

- **UserProfile**: For each user profile, we record the user's **name**, **reputation**, and unique **userID**. The **reputation** attribute measures the degree to which the community trusts a particular user. For example, a user can increase their reputation by posting helpful answers to questions.
- **Question**: For each question asked by a user, we record his or her **userID**, the question **title**, and unique **questionID**.
- **Answer**: We track all answers posted by users in the **Answer** table (see Table 1(c)). Each row stores the **userID** of the user that answered the question, the **questionID** of the question answered, and the **vote_count**, (i.e., the total number of users that voted that answer as being the most helpful). For example, the first row of Table 1(c) shows that user 'U4' (='Erika') answered question 'Q2' (='Inefficient database...') and, for her answer, she received a **vote_count** of 118.

Given this database instance, answer the following questions:

(a) [**2 points**] Which of the following is the meaning of the expression:

$$\sigma_{\texttt{reputation} \geq 200}(\texttt{UserProfile})$$

    1. It lists the **userID** and **name** of all users with a **reputation** greater than or equal to 200.
    2. It lists all **reputation**s that are greater than or equal to 200.
    3. It lists all tuples in the **UserProfile** table ((**userID**, **name**, and **reputation**) with a **reputation** greater than or equal to 200.
    4. None of the above. The real answer is . . . .

(b) [**5 points**] We want to list the **title**s of questions asked by "top" users, (i.e., users with a **reputation** greater than 75). Which, if any, of the following expressions achieve that? <u>Mark **all** valid expressions.</u>

    1. $\pi_{\texttt{title}}\big(\sigma_{\texttt{reputation}>75}(\texttt{UserProfile} \bowtie \texttt{Question}) \big)$

    2. $\sigma_{\texttt{reputation}>75}\big(\pi_{\texttt{title}}(\texttt{UserProfile} \bowtie \texttt{Question}) \big)$

    3. $\pi_{\texttt{title}}\big(\texttt{Question} \bowtie \sigma_{\texttt{reputation}>75}(\texttt{UserProfile}) \big)$

    4. $\pi_{\texttt{title}}\Big(\sigma_{\texttt{reputation}>75} \big(\texttt{UserProfile} \bowtie \pi_{\texttt{title,userID}}(\texttt{Question}) \big) \Big)$

    5. $\pi_{\texttt{title}}\Big(\sigma_{\texttt{reputation}>75}\big(\texttt{UserProfile} \bowtie \pi_{\texttt{title}}(\texttt{Question}) \big) \Big)$

(c) For the following expression:

$$\sigma_{\texttt{vote\_count}<25}(\texttt{Answer} \bowtie \texttt{Question})$$

    i. [**0 points**] *Optional*: describe in English what the expression does
    ii. [**1 point**] How many, and which are the columns (= attributes) in the answer?
    iii. [**3 points**] How many tuples are in the answer?

    iv. **[3 points]** List all the tuples in the answer, as a table.

(d) For the following expression:

$$\pi_{\texttt{userID,questionID}}(\texttt{Answer}) \div \pi_{\texttt{questionID}}\big(\sigma_{\texttt{userID='U4'}}(\texttt{Answer})\big)$$

    i. **[0 points]** *Optional*: describe in English what the expression does

    ii. **[2 points]** How many, and which are the columns (= attributes) in the answer?

    iii. **[3 points]** How many tuples are in the answer?

    iv. **[3 points]** List all the tuples in the answer, as a table.

(e) For the following expression:

$$\pi_{\texttt{C.userID}}\big(\rho_C(\texttt{UserProfile})\big) - \pi_{A.\texttt{userID}}\big($$
$$\rho_A(\texttt{Answer}) \bowtie_{A.\texttt{questionID}=B.\texttt{questionID} \wedge A.\texttt{vote\_count} < B.\texttt{vote\_count}} \rho_B(\texttt{Answer})\big)$$

    i. **[0 points]** *Optional*: describe in English what the expression does

    ii. **[2 points]** How many, and which are the columns (= attributes) in the answer?

    iii. **[3 points]** How many tuples are in the answer?

    iv. **[3 points]** List all the tuples in the answer, as a table.

Question 4: Relational Tuple Calculus (RTC) . . . . [15 points]

On separate page, with '[course-id] [hw#] [question#] [andrew-id] [your-name]'

We will again use the Q & A Website database from the last question (see Table 1).

(a) For the following RTC expression

$$\{t \mid \exists q \in \texttt{Question} \, (\, q.\texttt{userID} = \text{``}U5\text{''} \wedge q.\texttt{questionID} = t.\texttt{questionID} \,)\}$$

   i. [**0 points**] *Optional*: describe in English what the expression does
   ii. [**1 point**] How many, and which are the columns (= attributes) in the answer?
   iii. [**2 points**] How many tuples are in the answer?
   iv. [**2 points**] List all the tuples in the answer, as a table.

(b) For the following RTC expression

$$\{t \mid \exists p \in \texttt{UserProfile}, \exists a \in \texttt{Answer}$$
$$(\, p.\texttt{userID} = a.\texttt{userID}$$
$$\wedge \; p.\texttt{reputation} > 75$$
$$\wedge \; a.\texttt{vote\_count} > 100$$
$$\wedge \; t.\texttt{name} = p.\texttt{name}$$
$$\wedge \; t.\texttt{questionID} = a.\texttt{questionID})\}$$

   i. [**0 points**] *Optional*: describe in English what the expression does
   ii. [**1 point**] How many, and which are the columns (= attributes) in the answer?
   iii. [**2 points**] How many tuples are in the answer?
   iv. [**2 points**] List all the tuples in the answer, as a table.

(c) For the following RTC expression

$$\{t \mid \exists a1 \in \texttt{Answer}, \exists a2 \in \texttt{Answer}$$
$$(\, a1.\texttt{questionID} = a2.\texttt{questionID}$$
$$\wedge \, a1.\texttt{userID} > a2.\texttt{userID}$$
$$\wedge \, t.\texttt{userID1} = a1.\texttt{userID}$$
$$\wedge \, t.\texttt{userID2} = a2.\texttt{userID} \,)\}$$

   i. [**0 points**] *Optional*: describe in English what the expression does
   ii. [**1 point**] How many, and which are the columns (= attributes) in the answer?
   iii. [**2 points**] How many tuples are in the answer?
   iv. [**2 points**] Give, as a table, all of the tuples returned by the query.

Question 5: Relational Domain Calculus (RDC) ..... [15 points]

On separate page, with '[course-id] [hw#] [question#] [andrew-id] [your-name]'

We will again use the Q & A Website database from question 3 (see Table 1).

(a) For the following RDC expression
$$\{\langle u \rangle \mid \exists q, \exists v (\langle q, u, v \rangle \in \texttt{Answer} \ \wedge \ v < 50)\}$$

     i. [**0 points**]   *Optional*: describe in English what the expression does

     ii. [**1 point**]   How many, and which are the columns (= attributes) in the answer?

     iii. [**2 points**]   How many tuples are in the answer?

     iv. [**2 points**]   List all the tuples in the answer, as a table.

(b) For the following RDC expression:

$$\{\langle n1, n2 \rangle \mid \exists q, \exists u1, \exists v1, \exists n1, \exists r1, \exists u2, \exists v2, \exists n2, \exists r2 \ ($$
$$\langle q, u1, v1 \rangle \in \texttt{Answer}$$
$$\wedge \ \langle q, u2, v2 \rangle \in \texttt{Answer}$$
$$\wedge \ \langle u1, n1, r1 \rangle \in \texttt{UserProfile}$$
$$\wedge \ \langle u2, n2, r2 \rangle \in \texttt{UserProfile}$$
$$\wedge \ u1 > u2 \ )\}$$

     i. [**0 points**]   *Optional*: describe in English what the expression does

     ii. [**1 point**]   How many, and which are the columns (= attributes) in the answer?

     iii. [**2 points**]   How many tuples are in the answer?

     iv. [**2 points**]   List all the tuples in the answer, as a table.

(c) For the following RDC expression:

$$\left\{ \langle q, u1, v1 \rangle \ \middle| \ \left(\ \langle q, u1, v1 \rangle \in \texttt{Answer} \right.\right.$$
$$\left.\left. \wedge \ \forall u2 \left(\left(\exists v2 (\langle q, u2, v2 \rangle \in \texttt{Answer} \ ) \ \Rightarrow \ (v1 \geqslant v2)) \right) \right) \right) \right\}$$

     i. [**0 points**]   *Optional*: describe in English what the expression does

     ii. [**1 point**]   How many, and which are the columns (= attributes) in the answer?

     iii. [**2 points**]   How many tuples are in the answer?

     iv. [**2 points**]   List all the tuples in the answer, as a table.