CSE 344 Homework 3: SQL (advanced)

Objectives:
To practice advanced SQL. To get familiar with commercial database management systems (SQL Server) and to get familiar with using a database management system in the cloud (SQL Azure).

Assignment tools:
SQL Server on Windows Azure through SQL Azure.

Due date:
Thursday, Februray 6, 2014, at 11:59pm. Turn in your answers here.

What to turn in:
`hw3-queries.sql`

This homework is a continuation of homework 2, with two changes. The queries are more challenging, and you will have to use a commercial database system. SQLite simply cannot execute these queries in any reasonable amount of time; hence, we will use SQL Server, which has one of the most advanced query optimizers. SQL Server also has a very nice client application, SQL Server Management Studio, that you will get to use in this assignment.

Here is again the schema of the IMDB database, for your reference:

ACTOR (id, fname, lname, gender)
MOVIE (id, name, year)
DIRECTORS (id, fname, lname)
CASTS (pid, mid, role)
MOVIE_DIRECTORS (did, mid)
GENRE (mid, genre)

All `id` fields are integers. MOVIE.year is an integer. All other fields are character strings.

id column in **ACTOR**, **MOVIE** & **DIRECTOR** tables is a key for the respective table.

CASTS.pid refers to **ACTOR.id**
CASTS.mid refers to **MOVIE.id**
MOVIE_DIRECTORS.did refers to **DIRECTORS.id**
MOVIE_DIRECTORS.mid refers to **MOVIE.id**
GENRE.mid refers to **MOVIE.id**

In this homework, you will do three things. First, you will connect to a database system running as a service on Windows Azure. Second, you will write and test the six SQL queries below; keep in mind that the queries are quite challenging, both for you and for the database engine. Third, you will reflect on using a database management system running in a public cloud.

The good news is that the IMDB database is already uploaded on the server, and all indices are created; all you need is to do is to connect successfully, then run your queries.

A. Connecting to SQL Server on Windows Azure:

We use an instance of SQL Server running as a service in the Microsoft Azure Cloud. You can connect in two ways: (a) use your Web browser, https://m01rrgdwg2.database.windows.net (normally you will connect to the *IMDB* database, but the first time you should connect to the *master* database in order to change your password) or (b) run **SQL Server Management Studio 2008 R2** (it has to be R2 -- this is already installed on the PC labs and is also available on all VDI lab machines -- see: http://vdi.cs.washington.edu/vdi/),

In both cases, your login is your UW login (without '@washington.edu') and the initial password is given in class.

When connecting using SQL Server Management Studio, do the following

- Server type: Database Engine
- Server name: m01rrgdwg2.database.windows.net
- Authentication: SQL Server Authentication
- Login: Your UW netid
- Password: Initially, use the one we gave in class

Once you are connected, change your password by running the following command in the *master* database:

```
ALTER LOGIN yourlogin WITH PASSWORD='some_new_password' OLD_PASSWORD = 'old_password'
```

More precisely:

- If you connect using a Web browser, first connect to the *master* database, then run the ALTER LOGIN command above; then connect to the *IMDB* database.
- If you connect from Management Studio, then: in the Object Explorer on the left, select Databases -> System Databases -> master
  - Click on New Query (at the top)
  - Execute the ALTER LOGIN command above. Make sure to use some capital letters and some numbers.
  - To run queries against IMDB, **first** select the database called IMDB in the Object Explorer and **then** click on "New Query"

Note that you will get an error message if your new password is not sufficiently complex. If you have any problems connecting to Windows Azure, please let the instructor or the TAs know. Once you connect, for fun, try and run some of the queries from the previous homework and see how fast they run compared to SQLite.

B. SQL QUERIES (90 points; 15 points per question):

For each question below, write a single SQL query to answer that question. Add a comment to each query indicating the question number and the number of rows your query returns.

1. Consider all actors that had five or more roles in a movie in 2010. In homework 2, we asked you to list each such actor's name, the movie name, and the number of roles he/she played. Do the same thing, but instead of giving the number of roles, give the name of each role. Your answer should have one tuple for each combination of (actor, movie, role) - so if an actor has 10 roles in a given movie, there should be 10 tuples for that actor and movie. *Approx. 140 rows.*

2. For each year, count the number of movies in that year that had only female actors. Recall the meaning of the universal quantifier: a movie without any actors is also a movie with only female actors (since there are no male actors in such a movie!). *Approx. 130 rows.*

3. Now make a small change: for each year, report the percentage of movies with only female actors made that year, and also the total number of movies made that year. For example, one answer will be:

   ```
   1990 31.81 13522
   ```

   meaning that in 1990 there were 13,522 movies, and 31.81% had only female actors. You do not need to round your answer. *Approx. 130 rows.*

4. Find the film(s) with the largest cast. Return the movie title and the size of the cast. By "cast size" we mean the number of distinct actors that played in that movie: if an actor played multiple roles, or if the actor is simply listed more than once in CASTS, we still count her/him only once. You may *not* assume that only one film has the largest cast. *1 row. The cast size is around 1300.*

5. A decade is a sequence of 10 consecutive years. For example 1965, 1966, ..., 1974 is a decade, and so is 1967, 1968, ..., 1976. Find the decade with the largest number of films. *1 row. If you were to count those movies in the decade then you would get around 457,500.*

6. The Bacon number of an actor is the length of the shortest path between the actor and Kevin Bacon in the "co-acting" graph. That is, Kevin Bacon has Bacon number 0; all actors who acted in the same film as KB have Bacon number 1; all actors who acted in the same film as some actor with Bacon number 1 (but not with Bacon himself) have Bacon number 2, etc. Count how many actors have Bacon number is 2. *1 row. The number of actors is around 521,900*

C. Using a Cloud Service (10 points)

The DBMS that we use in this assignment is running somewhere in one of Microsoft's data centers. Comment on your experience using this DBMS cloud service. What do you think about the idea of offering a DBMS as a service in a public cloud?

Put all your code for part B (`SELECT-FROM-WHERE` code) in a file called `hw3-queries.sql` and add the answers to part C as SQL comments.