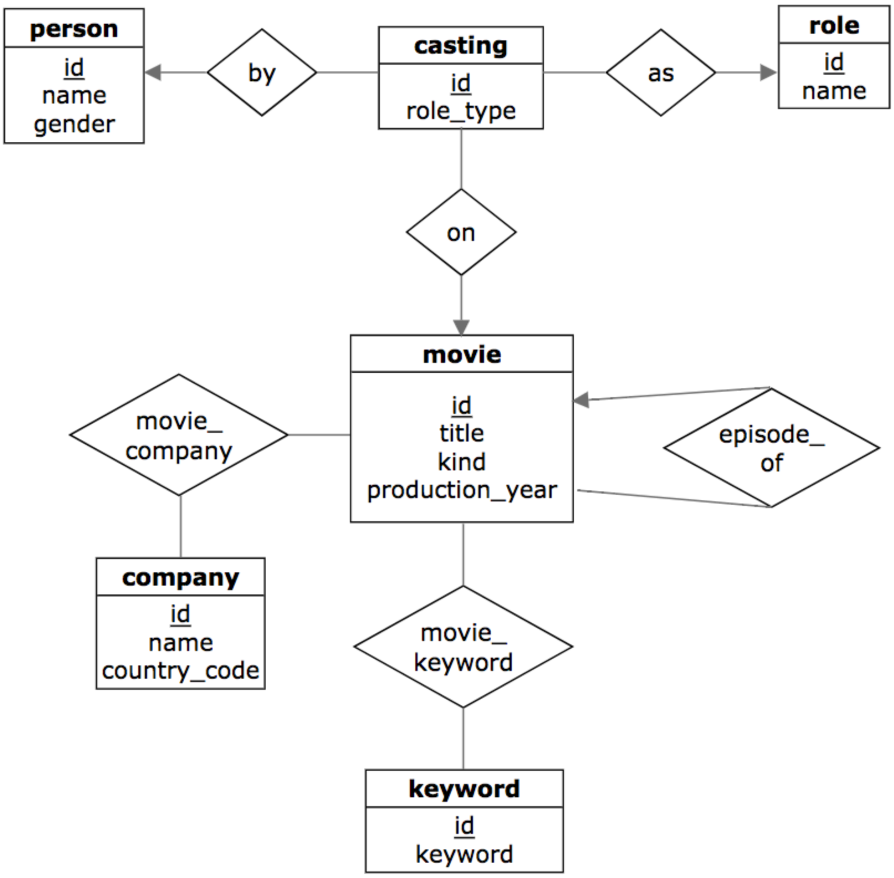
RAWDATA Assignment 1 – Querying IMDB with SQL

The main purpose of this assignment is to practice querying using SQL. The target is a database with a huge dump of "real world" data from IMDB covering more than 3.5 million movies. You can find and use this database from a remote server or download and install the database on your own computer. Check the instructions on page 4-5:

- *How to access a remote or a local database with the imdb data*
- *How to produce an output file and hand in your solution*

The available database, **imdb2,** is only an excerpt of what can be downloaded from IMDB and the tables are simplified and renamed. To get an overview of the database, consider the ER-diagram and the relational schema below.



Primary keys are underlined in the entities in the ER-diagram. The relational schema corresponding to the ER-diagram above (or as the DBC book puts is: reduced from the schema above) includes the following 8 tables.

person(id, name, gender)
casting(id, person_id, movie_id, role_id, role_type)
role(id, name)
movie(id, title, kind, production_year, episode_of_id)
company(id, name, country_code)
movie_company(movie_id, company_id)
keyword(id, keyword)
movie_keyword(movie_id, keyword_id)

Notice that the primary keys are shown (underlined) and the foreign keys (the latter not shown) for the relational schema can be inferred from the ER-diagram.

The database is quite large and the 8 tables have the following row counts.

Table name	**No of rows**
casting	49462688
role	3810797
company	302225
keyword	186008
movie_company	3505437
movie_keyword	6103575
person	5428172
movie	3570524

Try to study the content before you begin, for instance by running some simple queries. Notice

- Not only movies are included in the movie table (check the values for the "kind" attribute)
- Not only actors are included in the person table (check "role_type")

What to do

Write SQL queries to answer the questions below. Notice that query-answers are listed, so for each you will "simply" have to find an SQL-expression, that produces the given answer.
See instructions, by the end of this note, on how to hand in the result.

Important: Work individually, then work in groups, then hand-in by group

You are supposed to submit one hand in per group and we strongly recommend you to work with the questions individually as well as in groups. Discuss and decide on what you consider to be the best solutions and hand these in.

Questions

Observe: Most queries to answer the questions below should return in few seconds. Exceptions are e) and g). Expect slower response from these. However, stop your query evaluation if you experience significantly long waiting (several minutes) and reconsider your query expression. **Try to avoid leaving query processes running if you use the remote server**. Too many of these will slow everything down for everybody.

a) Find the titles and production year of all movies with a title starting with 'Pirates of the Caribbean'.

```
                         title                          | production_year
--------------------------------------------------------+-----------------
 Pirates of the Caribbean 6                             |
 Pirates of the Caribbean: At World's End               |            2007
 Pirates of the Caribbean: Dead Man's Chest             |            2006
 Pirates of the Caribbean: Dead Men Tell No Tales       |            2017
 Pirates of the Caribbean: On Stranger Tides            |            2011
 Pirates of the Caribbean: Tales of the Code: Wedlocked |            2011
 Pirates of the Caribbean: The Curse of the Black Pearl |            2003
 Pirates of the Caribbean: The Young Sparrow            |            2016
(8 rows)
```

b) How many movies were produced in 2004?

```
count
-------
 14143
(1 row)
```

c) Find the title of all video games with Mads Mikkelsen.

```
      title
-------------------
 Quantum of Solace
```

```
(1 row)
```

d) Find the different role_types that Kevin Bacon has had.

```
    role_type
-----------------
 actor
 cinematographer
 director
 editor
 producer
 writer
(6 rows)
```

e) Find for each role_type the number of persons casted with this role_type. Show in descending order by the number.

```
      role_type      |  count
---------------------+----------
 actor               | 16622527
 actress             |  9906721
 miscellaneous crew  |  6170099
 producer            |  5936394
 writer              |  3799109
 director            |  2353225
 editor              |  1599479
 cinematographer     |  1234113
 composer            |  1055730
 production designer |   421641
 costume designer    |   363650
(11 rows)
```

f) Find the title of movies directed by Ridley Scott from 2004. 2006, 2008 or 2010..

```
    title
--------------
 A Good Year
 Body of Lies
 Robin Hood
(3 rows)
```

g) What is the highest number of actors casted for one movie?

```
max
-----
 943
(1 row)
```

h) How many actors have acted together with Kevin Bacon?

```
count
-------
  5721
(1 row)
```

i) Find the title of all movies that are assigned to the keyword 'elephant-fears-mouse'

```
               title
------------------------------------
 Shooting Fish in a Barrel
 The Two Mrs. Nahasapeemapetilons
 Acrobatty Bunny
 Dumbo
 Goliath II
 Krazy Kat - Bugologist
 The Adventures of Baron Munchausen
 The Mite Makes Right
 Tweety's Circus
 Unnatural History
 Woodman, Spare That Tree
```

```
(11 rows)
```

j) Find the title of all movies that are assigned to keywords 'dancing' as well as to 'elephant-fears-mouse'.

```
               title
------------------------------------
 Acrobatty Bunny
 The Adventures of Baron Munchausen
(2 rows)
```

k) Find the title and production year of movies from "Paramount" in Sweden produced after 2004 ordered by title.

```
     title      | production_year
----------------+-----------------
 Choke          |            2008
 Dreamgirls     |            2006
 Hotelliggaren  |            2005
 Hustle & Flow  |            2005
 Madagascar     |            2005
 Over the Hedge |            2006
 Stardust       |            2007
(7 rows)
```

l)Find the title of movies that have casted to a role named "The Singing Kid"

```
title
--------
 Broken
(1 row)
```

m) Who have played a role named "Bilbo" in a movie where the word "Smaug" appear in the title.

```
      name       |                    title
-----------------+---------------------------------------------
 Baxter, Daniel  | How the Desolation of Smaug Should Have Ended
 Freeman, Martin | The Hobbit: The Desolation of Smaug
(2 rows)
```

How to access a remote or a local database with the imdb data

The data is stored in a database called **imdb2**. Choose either of the following to access this.
Method 1 (remote DB): You can find the **imdb2** database on the server **rawdata.ruc.dk** (user "guest" and password "guest_").
Method 2 (local DB): You can alternatively download and install the database on your own computer. Do the following:

1) Download the file **imdb2.backup.zip** using the link on Moodle (more than 500 MB).
2) Unzip to get the file **imdb2.backup** (maybe around 2 GB).
3) Open your command line interface and change to the directory where you placed the unzipped file **imdb2.backup**.
4) Run the two commands (leading to a 6-10 GB database in your Postgres DBMS)

```
psql -U postgres -c "create database imdb2"
psql -U postgres -d imdb2 -f imdb2.backup
```

5) When finished using the imdb2 database for Assignment 1 you may consider (to save space) to delete the two files and to drop the database. To drop the database use this command:

```
psql -U postgres -c "drop database imdb2"
```

The remote server will probably be slower that your own DB server, but it shouldn't be a problem for this assignment.

How to produce an output file and hand in your solution

When you have tested all your solutions to questions a) to m) one by one, include all these in a single SQL script file **assignment1.sql** with content like:

```
-- GROUP: <group-name>, MEMBERS: <name1>, <name2>, ...
-- a)
SELECT ...
...

-- b)
SELECT ...
...

-- c)
SELECT ...
...
...
```

To indicate "no solution" for a question, just write "-- no solution" in place of the SQL expression. To include timings in the output (not required, not shown above) you can add a first line: "\timing".

To hand in your solution do the following:

- Generate an output file "xxxx-assignment1.txt" (xxxx is your groups name) by running the command (if remote DB):

  ```
  psql  -h rawdata.ruc.dk -p 5432 -U guest -a -d imdb2 -f assignment1.sql >
  xxxx-assignment1.txt
  ```

  or (if local DB):

  ```
  psql -U postgres -a -d imdb2 -f assignment1.sql > xxxx-assignment1.txt
  ```

  What's going on here is that all the SQL-code in your input file **assignment1.sql** will be processed and the output will be written to the file **xxxx-assignment1.txt** including echoes of the SQL expressions and the -- comments (the -a option takes care of the echoing).
- Check your output file **xxxx-assignment1.txt** and hand it in on the Moodle page.
  OBS: Just one hand-in per group

Main reference for this assignment is:

- [DSC] Database System Concepts, Abraham Silberschatz, Henry Korth, S. Sudarshan, 7th Edition, 2019, chapter 3 and 4

To read more about how to use the command line tool **psql** consider:

- PostgreSQL Tutorial, 17 Practical psql Commands That You Don't Want To Miss http://www.postgresqltutorial.com/psql-commands/
- [PGMAN] (html) psql chapter https://www.postgresql.org/docs/current/static/app-psql.html (same as [PGMAN] (pdf) 1793-1829)