DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Databases Project – Spring 2015

In this project the students have to design a database schema and application which analyzes and maintains databases about movies based on the IMDB database. Below you will find a detailed description of the tasks to be carried out throughout the project.

Deliverable 1: Create ER model. Design and Create Schema.

(Due: 22nd March 2015)

The students will use the data that will find in the following IMDB data files:

- Person
- Alternative_name
- Production
- Alternative_title
- Company
- Production_cast
- Production_company
- Character

The goal of this deliverable is to design an ER model, a corresponding relational schema and create the database tables in the given database. The organization of the data in files and the given description ***does not imply*** neither an ER model nor a relational schema. It is given to help the student understand the format of the data faster. Finally, a discussion about constraints and removing redundant information should be included in the project report.

In the 1st deliverable the students should:

1. Create the ER model for these data.
2. Design the database and the constraints needed to maintain the database consistent.
3. Create the SQL commands to create the tables in Oracle. (Provide the SQL commands)
4. Describe their work in the form of a report which should contain an ER diagram, SQL DDL code for table creation, description of the data constraints and justification of the design choices (in a few paragraphs). The report should be submitted as a single pdf file (one pdf per group).

DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Deliverable 2: Import Data. Basic SQL queries.

(Due: 26nd April 2015)

The students should accommodate the insertion of new data are inserted in any table. Moreover, they should implement a simple query which can search for a keyword in any table. The user should be able to see more details of the result of the query (e.g., if someone searches for Robert De Niro and the result is a series of movies, then he/she should be able to see more details for some movies – through a hyperlink for example). A few more queries should be implemented:

a) Compute the number of movies per year. Make sure to include tv and video movies.
b) Compute the ten countries with most production companies.
c) Compute the min, max and average career duration.
(A career length is implied by the first and last production of a person)
d) Compute the min, max and average number of actors in a production.
e) Compute the min, max and average height of female persons.
f) List all pairs of persons and movies where the person has both directed the movie and acted in the movie. Do not include tv and video movies.
g) List the three most popular character names.

In the 2nd deliverable the students should:

1. Parse the given IMDB data and import them in the created database as described in your 1st deliverable.
2. Implement (with SQL) the simple search queries and the follow-up search queries of the result of the initial search.
3. Implement the queries described above.
4. Build an interface to access and visualize the data. Website or a java application are good choices, but students are free to choose any technology they want.
5. Extend the project report from the first deliverable with the description of the work done for the second deliverable and an explanation for the design choices. Include any changes to the design covered in the first deliverable with justification of the changes. Include the screenshot of the interface and the description of the way search functionality is implemented in the application. The report should be submitted as a single pdf file.

DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Deliverable 3: Interesting SQL queries.

(Due: 31st May 2015)

A series of more interesting queries should be implemented with SQL and/or using the preferred application programming language.

a) Find the actors and actresses (and report the productions) who played in a production where they were 55 or more year older than the youngest actor/actress playing.
b) Given an actor, compute his most productive year.
c) Given a year, list the company with the highest number of productions in each genre.
d) Compute who worked with spouses/children/potential relatives on the same production. (You can assume that the same *real surname* implies a relation)
e) Compute the of average number of actors per production per year
f) Compute the average number of episodes per season.
g) Compute the average number of seasons per series.
h) Compute the top ten tv-series (by number of seasons).
i) Compute the top ten tv-series (by number of episodes per season).
j) Find actors, actresses and directors who have movies (including tv movies and video movies) released after their death.
k) For each year, show three companies that released the most movies.
l) List all living people who are opera singers ordered from youngest to oldest.
m) List 10 most ambiguous credits (pairs of people and productions) ordered by the degree of ambiguity. A credit is ambiguous if either a person has multiple alternative names or a production has multiple alternative titles. The degree of ambiguity is a product of the number of possible names (real name + all alternatives) and the number of possible titles (real + alternatives).
n) For each country, list the most frequent character name that appears in the productions of a production company (not a distributor) from that country.

In the 3rd deliverable the students should:

1. Accommodate all above queries by giving the corresponding SQL code.
2. Explain the necessities of indexes based on the queries and the query plans that you can find from the system (you are free to select any 3 queries you like from the queries of the 3rd deliverable).
3. Report the running time of all queries in milliseconds and explain the distribution of the cost (based again on the plans) for 3 queries selected in part 2.
4. Visualize the results of the queries (in case they are not scalar).

DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

5. Build an interface to run queries/insert data/delete data giving as parameters the details of the queries.
6. Complete the project report written for the previous deliverables by adding description of the queries and interfaces, explanation for the design choices, analysis of the chosen queries, as well as the changes compared to the work described in the previous deliverables. The report should be submitted as a single pdf file.

You can find the data here: http://diaswww.epfl.ch/courses/db2015/project/Movies.tar.gz

DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

IMDB data description

Person

This file describes the people that are involved in making movies.

1. id
2. Name
   - The full name and surname of a person working for a production.
3. Gender
4. Trivia
   - Short trivia considering the person.
5. Quotes
   - Quotes of the person from a movie or interview.
6. Birth date
7. Death date
8. Birth name
9. Mini biography
   - A short biography of the person.
10. Spouse
    - The name of the spouse of the person.
11. Height
    - The height of the person (can be either in cm or inches)

Alternative_name

This file describes any alternative names a person may have.

1. id
2. person_id
3. Name

Production

This file describes any production of movie, series episode or series in general.

1. id
2. title
   - The title of the production.

DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

3. production_year
4. series_id
   - The id of a series that an episode may belong to.
5. season_number
   - The number of the season of a series that an episode may belong to.
6. episode_number
   - The number of the episode in a season that an episode may belong to.
7. series_years
   - The range of years that a series has been aired.
8. Kind
   - What kind of production this is (tv_series, tv_movie, episode, movie, video movie, video game).
9. Genre
   - What genre of filmmaking is this production (Action, Horror etc.).

Alternative_title

This file describes any alternative title a production may have.

1. id
2. production_id
3. title

Character

This file describes a character that may be playing a role in a production

1. id
2. name

Production_cast

This file provides the connection between the people, productions and characters in a given production.

1. production_id
2. person_id
3. character_id
4. role
   - The role of a person in a production (writer, director, actor etc.).

DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Company

This file describes companies involved in the industry.

1. id
2. country_code
3. name

Production_company

This file describes the companies that participate in the creation of the production.

1. id
2. company_id
3. production_id
4. kind
    - The type of the company (distributors or production company)