**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/

# Databases Project – Spring 2013

In this project the students have to design a database schema and application which analyzes and maintains a database with statistics about the Olympic Games. 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.*

The students will use the data supplied by the course staff. The data is contained in 8 files in comma separated format. The format of these files is briefly described in the Appendix.

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 is expected.

In the 1st deliverable the students should:

1. Create the ER model for the 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.
4. Describe their work in the form of a report which should contain ER diagram, SQL DDL code for table creation, description of the data constraints and justification of the design choices in text format.

## *Deliverable 2: Import Data. Basic SQL queries.*

The students should accommodate the situation where new data is 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 information on Vreni Schneider's medals at the Olympics and the result has multiple Olympics games, he/she should be able to see statistics for individual games  – for example, through a hyperlink). A few more queries should be implemented:

A. Print the names of athletes who won medals at both summer and winter Olympics.
B. Print the names of gold medalists in sports which appeared only once at the Olympics.
C. For each country print the place where it won its first medal.
D. Print the name of the country which won the most medals in summer Olympics and the country which won the most medals in winter Olympics.
E. List all cities which hosted the Olympics more than once.
F. List names of all athletes who competed for more than one country.
G. For each Olympic Games print the name of the country with the most participants.
H. List all countries which didn't ever win a medal.

In the 2nd deliverable the students should:

**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/

1. Import the data from the given files into the created database.
2. Accommodate the import of new data in the database they created in the 1$^{st}$ deliverable.
3. Implement (using SQL) the simple search queries and the follow-up search queries of the result of the initial search.
4. Implement the queries described above.
5. Build an interface to access and visualize the data. A web front-end is proposed.
6. Extend the project report from the first deliverable with the description of the work done for the second deliverable and rationale for the design choices. Include any changes to the design covered in the first deliverable with justification of the changes.

## *Deliverable 3: Interesting SQL queries.*

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

I.   Compute medal table for the specific Olympic Games supplied by the user. Medal table should contain country's IOC code followed by the number of gold, silver, bronze and total medals. It should first be sorted by the number of gold, then silvers and finally bronzes.
J.   For each sport, list the 3 nations which have won the most medals
K.   Compute which country in which Olympics has benefited the most from playing in front of the home crowd. The benefit is computed as the number of places it has advanced its position on the medal table compared to its average position for all Olympic Games. Repeat this computation separately for winter and summer games.
L.   List top 10 nations according to their success in team sports. Use average number of medalists for each medal awarded to a particular nation.
M.   List all Olympians who won medals for multiple nations.
N.   List all nations whose first medal was gold, all nations whose first medal was silver and all nations whose first medal was bronze. If nation won more than one medal at the first Olympics it won a medal, consider that it won the "shinier" medal first. For example if a country didn't win any medals before games in 1960 and then it won a gold and a bronze, then its first medal is a gold.
O.   For all disciplines, compute the country which waited the most between two successive medals.
P.   List all events for which all medals are won by athletes from the same country
Q.   For each Olympic Games, list the name of the country which scored the largest percentage of the medals.
R.   For all individual sports, compute the most top 10 countries according to their success score. Success score of a country is sum of success points of all its medalists: gold medal is worth 3 points, silver 2 points, and bronze 1 point. Shared medal is worth half the points of the non-shared medal.
S.   List names of all athletes who won medals both in individual and team sports.
T.   List names of all athletes who won gold in team sports, but only won silvers or bronzes individually.
U.   List names of all events and Olympic Games for which the individual or team has defended a title from the previous games.
V.   List top 10 countries according to their success on the events which appear at the Olympics for the first time. Present the list in the form of the medal table (as described for query I).

**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/

In the 3<sup>rd</sup> 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 3<sup>rd</sup> deliverable).
3. Report the performance of all queries and explain the distribution of the cost for 3 queries selected in part 2.
4. Visualize the results of the queries (in case they are not scalar create a table where the results will be clear).
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, rationale behind design choices, analysis of the chosen queries, as well as changes compared to the work described in the previous deliverables

**You can find the data here:  http://diaswww.epfl.ch/courses/db2013/project**

**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/

# Olympics data description

**Games.csv**
- Name, number_of_countries, number_of_athletes, number_of_events, host_city, host_country

**Countries.csv**
- Name, ioc_code

**Sports.csv**
- Name

**Disciplines.csv**
- Name, sport

**Events.csv**
- Name, discipline, games

**Medals.csv**
- Country, event, medal, medalist names, olympics

**Participants.csv**
- Name, country, olympics, sport

**Athletes.csv**
- Name

Note: some values might be missing since data is incomplete. Feel free to add missing data.
All data in these files is in the form of string, however many queries could be expressed easier using attributes with numeric or categorical values. We encourage you to consider transforming some fields into multiple fields and adding fields.