**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 2012

In this project the students have to design a database schema and application which analyzes and maintains a database with statistics about NBA players and coaches. 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 contained in the following data files:

- coaches_career.csv
- coaches_data.csv
- draft.csv
- player_allstar.csv
- player_carrer.csv
- player_playoffs.csv
- player_playoffs_career.csv
- player_regular_season.csv
- players.csv
- team_season.csv
- teams.csv

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.

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

The students should accommodate the situation where new data is inserted in any table. Moreover, a simple query which can search for a keyword in any table should be implemented. The user should be able to see more details of the result of the query (e.g., if someone searches for Michael Jordan's regular season statistics and the result has multiple seasons, he/she should be able to see statistics for individual seasons – for example, through a hyperlink). A few more queries should be implemented:

**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

a) Print the last and first name of players/coaches who participated in NBA both as a player and as a coach.
b) Print the last and first name of those who participated in NBA as both a player and a coach in the same season.
c) Print the name of the school with the highest number of players sent to the NBA.
d) Print the names of coaches who participated in both leagues (NBA and ABA).
e) Compute the highest scoring and lowest scoring player for each season.
f) Print the names of oldest and youngest player that have participated in the playoffs for each season.

In the 2nd deliverable the students should:

1. Import the data from the given csv files into the created database.
2. Accommodate the import of new data in the database they created in the 1st 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.

## *Deliverable 3: Interesting SQL queries.*

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

a) List the name of the schools according to the number of players they sent to the NBA. Sort them in descending order by number of drafted players.
b) List the name of the schools according to the number of players they sent to the ABA. Sort them in descending order by number of drafted players.
c) List the average weight, average height and average age, of teams of coaches with more than XXX season career wins and more than YYY win percentage, in each season they coached. (XXX and YYY are parameters. Try with combinations: {XXX,YYY}={<1000,70%>,<1000,60%>,<1000,50%>,<700,55%>,<700,45%>}. Sort the result by year in ascending order.
d) List the last and first name of the players which have more than 12,000 rebounds and are shorter than the average height of players who have at least 10,000 rebounds (if any).
e) List the last and first name of the players who played for a Chicago team and Houston team.
f) List the top 20 career scorers of NBA.
g) For coaches who coached at most 7 seasons but more than 1 season, who are the three more successful? (Success rate is season win percentage: season_win / (season_win + season_loss)). Be sure to count all seasons when computing the percentage.
h) List the last and first names of the top 30 TENDEX players, ordered by descending TENDEX value (Use season stats). (TENDEX=(points+reb+ass+st+blk-missedFT-missedFG-TO)/minutes)
i) List the last and first names of the top 10 TENDEX players, ordered by descending TENDEX value (Use playoff stats). (TENDEX=(points+reb+ass+st+blk-missedFT-missedFG-TO)/minutes)

**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

j) Compute the least successful draft year – the year when the largest percentage of drafted players never played in any of the leagues.
k) Compute the best teams according to statistics: for each season and for each team compute TENDEX values for its best 5 players. Sum these values for each team to compute TEAM TENDEX value. For each season list the team with the best win/loss percentage and the team with the highest TEAM TENDEX value.
l) List the best 10 schools for each of the following categories: scorers, rebounders, blockers. Each school's category ranking is computed as the average of the statistical value for 5 best players that went to that school. Use player's career average for inputs.
m) Compute which was the team with most wins in regular season during which it changed 2, 3 and 4 coaches.
n) List all players which never played for the team that drafted them.

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 4 queries you like from the queries of the 3rd deliverable).
3. Report the performance of all queries and explain the distribution of the cost (based again on the plans of the same 4 queries as above).
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.

**You can find the data here: http://diaswww.epfl.ch/courses/db2012/NBA/**

**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

# NBA data description

## Coaches data

**Career data**

- CoachID, First name, Last name, Season wins, Season losses, Playoff wins, Playoff losses

**Season data**

- CoachID, Year, YearOrder, First name, Last name, Season wins, Season losses, Playoff wins, Playoff losses, Team

*Comments:*

- YearOrder field is used in cases when multiple coaches have coached single team during a season

## Draft data

**Draft data**

- DraftYear, DraftRound, Selection, Team, First name, Last name, PlayerID, DraftedFrom, League

*Comments:*

- PlayerID is an optional field which has value only for players which played in any of the leagues
- DraftedFrom field contains the name of the college or high school for players that played in the US, or country of the origin or the name of the previous team for international players

## Player data

**Player data**

- PlayerID, First name, Last name, Position, First Season, Last Season, HeightFeet, HeightInches, Weight, College, Birthdata

**AllStar games data**

- PlayerID, Year, First name, Last name, Conference, League, GP, minutes, pts, dreb, oreb, reb, asts, stl, blk, turnover, pf, fga, fgm, fta, ftm, tpa, tpm

**Regular season data**

- PlayerID, Year, First name, Last name, Team, League, GP, minutes, pts, dreb, oreb, reb, asts, stl, blk, turnover, pf, fga, fgm, fta, ftm, tpa, tpm

**Playoff data**

- PlayerID, Year, First name, Last name, Team, League, GP, minutes, pts, dreb, oreb, reb, asts,

**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/

stl, blk, turnover, pf, fga, fgm, fta, ftm, tpa, tpm

**Career regular season data**

- PlayerID, First name, Last name, League, GP, minutes, pts, dreb, oreb, reb, asts, stl, blk, turnover, pf, fga, fgm, fta, ftm, tpa, tpm

**Career playoff data**

- PlayerID, First name, Last name, League, GP, minutes, pts, dreb, oreb, reb, asts, stl, blk, turnover, pf, fga, fgm, fta, ftm, tpa, tpm

*Comments:*

- Some statistics can be shown as NULL or 0 when their values is 0, for example when player didn't have any blocks in any of the playoff games
- In the All Star file, GP which stands for games played is always 1, because there is only a single All Star game each season
- Some statistics, such as turnovers, are not available for the early years

# Team data

**Team data**

- TeamID, Location, Name, League

**Season data**

- TeamID, Year, League, o_fgm, o_fga, o_ftm, o_fta, o_oreb, o_dreb, o_reb, o_asts, o_pf, o_stl, o_to, o_blk, o_3pm, o_3pa, o_pts, d_fgm, d_fga, d_ftm, d_fta, d_oreb, d_dreb, d_reb, d_asts, d_pf, d_stl, d_to, d_blk, d_3pm, d_3pa, d_pts, Pace, Won, Lost

*Comments:*

- Statistics for the more recent years are split into offence (labels starting with o_) and defence (labeled d_) statistics, for the early years, they are combined and reported in the offence category
- Some statistics were not kept for the early seasons, for those, all values are 0