Advanced Databases 2015 **Dr. Pierpaolo Dondio**

Assignment – Relational and Dimensional Model, ETL

(20% of final exam marks)

Due Date: Monday 4th January (midnight)

In this assignment you are required to create a relational database, populate it, create a dimensional model and perform an ETL process at two different points in time.

Your dataset is about players' statistics during the 2011/2012 Premier League season. For each player, the dataset contains match-level stats such as number of passes, goals, shots...

The relational database you are asked to create must have the following schema:

Table Player

Player_ID	Pl_name	Pl_surname	TeamID (FK)

Table Stadium

Stadium_ID	St_name	City	Capacity	TeamID (FK)

Table Teams

Team_ID	Team_name	YearOfFound

Table Matches

Team_A	Team_B	M_date

Table Players_stats

Team_A_ID	Team_B_ID	M_date	Player_ID (FK)	MinPlayed	Goals	Shot_on	Shot_off	Penalty	Pass OK	Pass KO
Composite FK to table Matches										

The data to be loaded into this relational DB are in the script insert.sql and in a csv file (premier.csv).

The script Insert.sql populates the table Teams and the table Stadiums.

The csv file has the following format:

m_date,player_id,pl_surname,pl_name,team_name,team_id,opposition_name,opposition_id, min_played,shot_on,shot_off,penalty,pass_ok,pass_ko

The meaning of each field is as follows:
Goal = number of goals scored during the match
m_date = date of the match
player_id = unique id of the player
pl_surname = surname of the player
pl_name = name of the player
team_name = name of the player's team
team_id = id of the player's team
opposition_name = name of the opposing team (the one where the player is not playing in)
opposition_id = id of the opposing team (the one where the player is not playing in)
min_played = number of minutes played by the player in the match
shot_on = number of shots on target by the player
shot_off = number of shots off target by the player
penalty = number of goals scored by the player with a penalty
pass_ok = number of passes completed by the player
pass_ko = number of unsuccessful passes by the player

Note that the team names and team ids used in the sql script and in the csv file do match!

In the rest of the assignment, assume the following:
1. Player name + team name are unique in all the database, players do not change team during the season!
2. A team has only one home stadium

Deliverables

a. You are required to create the above relational model using Oracle or MySQL. Remember to implement the primary and foreign key constraints.
b. You are then required to populate the data by loading the data from the sql script and from the csv into the relational model. Suggestion: you could load all the csv into a big denormalized table, and then move the data into the appropriate DB tables.
c. Once the database has been loaded, you are required to perform an initial ETL process to move data into the data warehouse. You are required to define a sql script (that you might bundle into a procedure(s) or/and triggers) that automatically executes the process, creating the required stage tables and all the data matching procedures.

d. Finally, you are required to perform a second ETL using the additional data contained in the etl2.csv files. This file contains a number of new records describing new statistics for new and old players (but no new matches). First you need to load the data into some staging table and then perform the second ETL reusing some of the stage tables and procedures used for the first ETL.
e. Provide 2 sample queries that could be executed over the dimensional model to create reports about teams and players.
f. Document all your work in a document containing the commented script and a detailed explanation of how your code works, your design choices, the instructions about how to run your code and some evidence/test that the script is working correctly. The scripts have to run in ORACLE or MYSQL! You will demo your work on week 13.

IMPORTANT:
The data warehouse has the following dimensional model:

Fact Table
Fact_Stats(**date_sk**,**player_sk**,**team_sk**,**opponent_sk**,**stadium_sk**,min_played,goals,shot_on,shot_off,penalty,pass_ok,pass_ko)

Dimension Player
DimPlayer(**player_sk**,player_name,player_surname,player_age)

Dimension Team (used for the team and the opponent)
DimTeam(**team_sk**, team_name, year_of_foundation)

Dimension Time
DimTime(**date_sk**,year,month,day)

Dimension Stadium
DimStadium(**stadium_sk**,stadium_name,stadium_city,capacity)