Homework #4 **Due: turned in by Wed 10/18/2017 before class**

Wenbo Liu

(put your name above)

Total grade: _____ out of ___100___ points

There are 3 numbered questions. Please answer them all and submit your assignment as a single PDF file by uploading it to the HW4 drop-box on the course website.

For the first three questions, be sure to properly cite the source of reference. See the following instructions for citation style (https://www.library.cornell.edu/research/citation/apa). Basic examples:

Reference citations in text:

as has been shown (Leiter & Maslach, 1998)	-- with authors
on climate change (weather.com, 1997)	-- without authors

List of references at the end (also known as bibliography):

- Arrington, M. (2008, August 5). The viral video guy gets $1 million in funding. http://techcrunch.com/2008/08/05/the-viral-video-guy-gets-1-million-in-funding/

- U.S. Department of Health and Human Services. (2005). Medicaid drug price comparisons: Average manufacturer price to published prices (OIG publication No. OEI-05-05- 00240). Retrieved from http://www.oig.hhs.gov/oei/reports/oei-05-05-00240.pdf

1. Concepts
In your own words define the following terms AND describe the relationship of each term to other term(s) in the list:

A. ERP
B. Database
C. Data warehouse
D. Data mart
E. OLAP
F. OLTP
G. Data Mining
H. Business Intelligence

Provide your answers in a concise way within one or two pages (not including bibliography).

- A: ERP: Stands for Enterprise resource planning. It uses databases maintained by a DBMS to manage records of core business activities from a set of integrated applications, keeping track of business resources and transactions. As data become not current, it can be stored in a data warehouse.

- B: Databases: A database can be defined as a collection of data that has been organized for storage and accessibility, keeping track transactions/records. A traditional database is usually optimized for OLTP: data is usually current, and emphasis on speed of processing short-queries (Insert, delete, update) on single transactions/records. A typical database uses a entity-model.

- C: Data Warehouse is a subset of database. A data warehouse is usually non-volatile(read-only), it stores historical information once that information within some database is no longer current or is not longer urgent for immediate uses, and it provides access to data in order to perform analysis and ad hoc queries. Data warehouses' depth can be enterprise-wide and can acquire information from several different

databases, therefore it usually has larger volume compared to other databases. They are usually optimized for OLAP: to execute complex queries and analyze large volume of data to support data-driven decisions.

- D: Data Mart can be considered as a subset of a data warehouse: it has the same functionalities of a data warehouse except it focus on a specific subset of business/single business line.
- E: OLAP stands for online analytical processing. As explained above, it is characterized by small number of complex queries that usually involve aggregating large number of rows of data. Its purpose is to analyze large amount of historical data and support data-driven decisions, and is usually applied on top of a dimensional schema.

- F: OLTP stands for online transactional processing, processes such as update, insert, and delete that keep track of business transactions/records. A single query might involve smaller amount of data, but the number of queries can become very large depending on the business process, therefore OLTP databases focus on numbers of queries executed within a unit time.

- G: Data Mining is the process of retrieving knowledge from data using statistical and machine learning techniques. Once a data warehouse is ready to use, we can apply data mining to data retrieved through OLAP from the warehouse.

- H: Business Intelligence is a broad aspect, it is the concept of utilizing data to support decisions and it involves data retrieval and aggregation, data analysis and visualization. In terms of enterprise analytical architecture, BI is one layer on top of data warehouse/data marts; it receives queries/command from the user, ask for related data from data warehouse/database, perform analysis on it and return the result/findings to the user in different forms (report, visualization, etc).

Citation

Elmasri, R. A., Navathe, S. B., & Castano, S. (2015). Fundamentals of database systems. S.l.: Pearson Addison Wesley.

2. Please provide short answers to the following questions:

a. What are the major differences between normalized ER Modeling and dimensional modeling (star schema)? (List at least three).

1: In a dimensional model, there are two distinct types of tables: fact tables and dimensional tables, and each of them store different information: fact tables store measurements and foreign keys referencing dimension tables; dimension tables store attributes of each dimension. In a normalized ER model, however, there usually isn't a single table dedicated to store facts or attributes.

2: A dimensional model is de-normalized; Depending on the defined grain of a dimensional model, some information within rows of a fact table can be somewhat redundant compared to a normalized ER model. In the expense of more disk space, a dimensional model is more efficient on retrieving and aggregating data by avoiding joining multiple tables.

3: An ER model is usually deployed on a OLTP database, whereas a dimensional model is usually deployed on a OLAP focused data warehouse.

4:ER modeled database typically use current data, whereas DM database uses historical data.

b. What are the main reasons to use dimensional modeling instead of normalized ER modeling for data warehousing designs? (List at least two).

1: As stated above, DM is more efficient on retrieving/aggregating large amount of data because if avoids joining multiple tables. Since DW usually has larger volume and it is usually used to perform analysis on large amount of data, DM is more efficient.

2: Since DW typically acquire data from multiple data sources, it should be more resilient to design changes/new data types compared to traditional database. DM has more advantage over normalized ER modeling in this aspect because when we add new designs/data types into a DM, nothing existing needs to be changed; Adding new dimensions only require addition of one column to the fact table, and adding attributes doesn't affect fact table at all.

c. Explain the following concepts in a sentence or two.

1. Fact

Facts are measurements taken from a business process; they must be true to the pre-defined grain.

2. Grain

Grain is the 'unit' of a single row within a fact table. It defines levels of details of each row, and it specifies what each row represents within a fact table.

3. OLAP cube

An OLAP cube is a subset of the entire dataset sliced by some specific dimensions defined by a specific business process. By managing data using OLAP cubes, analysis and data retrieval becomes faster**.**

4. Snowflake schema

It is a variation of star-schema; unlike a star-schema where one dimension is related to one table, dimensional tables in a Snowflake schema are heretical by normalization; therefore one dimension in a fact table can have more than one dimension table.

3. The goal of this homework is to create a data warehouse star schema for tracking fantasy basketball. Fantasy basketball is a popular game for basketball fans. Here are some useful details:

- **Groups of users form a fantasy basketball league. Each league has an owner who is the creator of the league.**
- **A Fantasy League consists of a group of 6-12 Fantasy Teams (hence 6-12 users) who agree to play against each other.**
- **Each member user of a league operates a fantasy team.**
- **Each Fantasy Team consists of a number or real-life basketball players. At the beginning of the season, each user selects the real-life players that will be on his/her team during the Draft. Typically, a real-life player can only be on one Fantasy Team within a Fantasy League.**
- **Users can trade players with other Fantasy Teams to improve their team.**
- **The real-life statistics accumulated by the players on a team are aggregated and ranked against the same statistics for the other teams in the league. For example, in a league of 10 teams, the team the most rebounds over the season to date would be rewarded 10, the second highest gets 9 and so on.**
- **In fantasy basketball, a season may last the whole real-life basketball season. But there are also short formats such as a daily contest (which we do not model).**

Review the source data in the appendix. We will build a data warehouse from the source data to answer questions such as

- **Who are the most drafted players across all leagues?**
- **Which user has the highest number of assists in the current season? (it means the user's players' assists while the user has them).**
- **Who are the most traded players in a particular fantasy league?**
- **How are teams ranked in a league in terms of overall fantasy points (which can be calculated from the number of points, assists, rebounds, etc.)?**

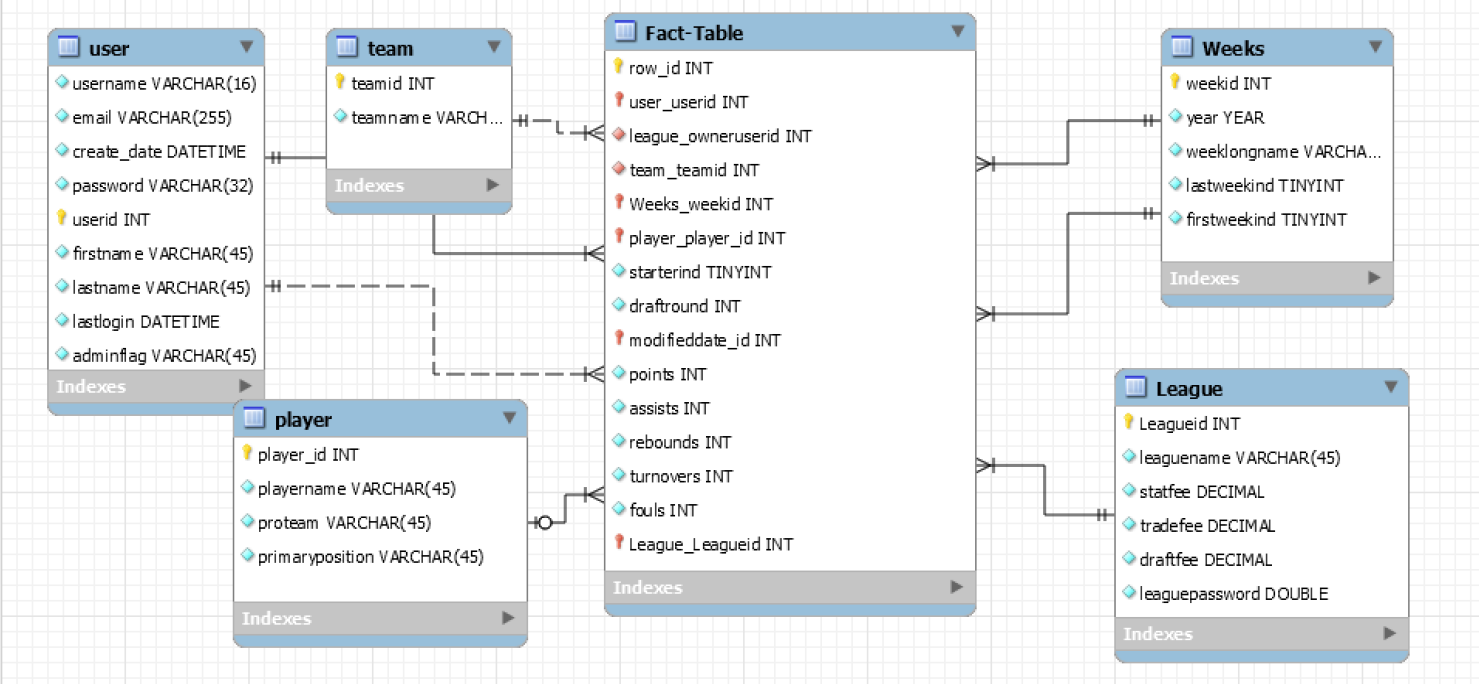
You can follow the following steps to build the data warehouse:

- **Step 1: What is the grain of the business process that we will model?**
- **Step 2: What are the facts?**
- **Step 3: What are the dimensions?**
- **Step 4: (Use MySQL Workbench) Draw an ER diagram with the fact and dimensions table. Identify the primary and foreign keys.**

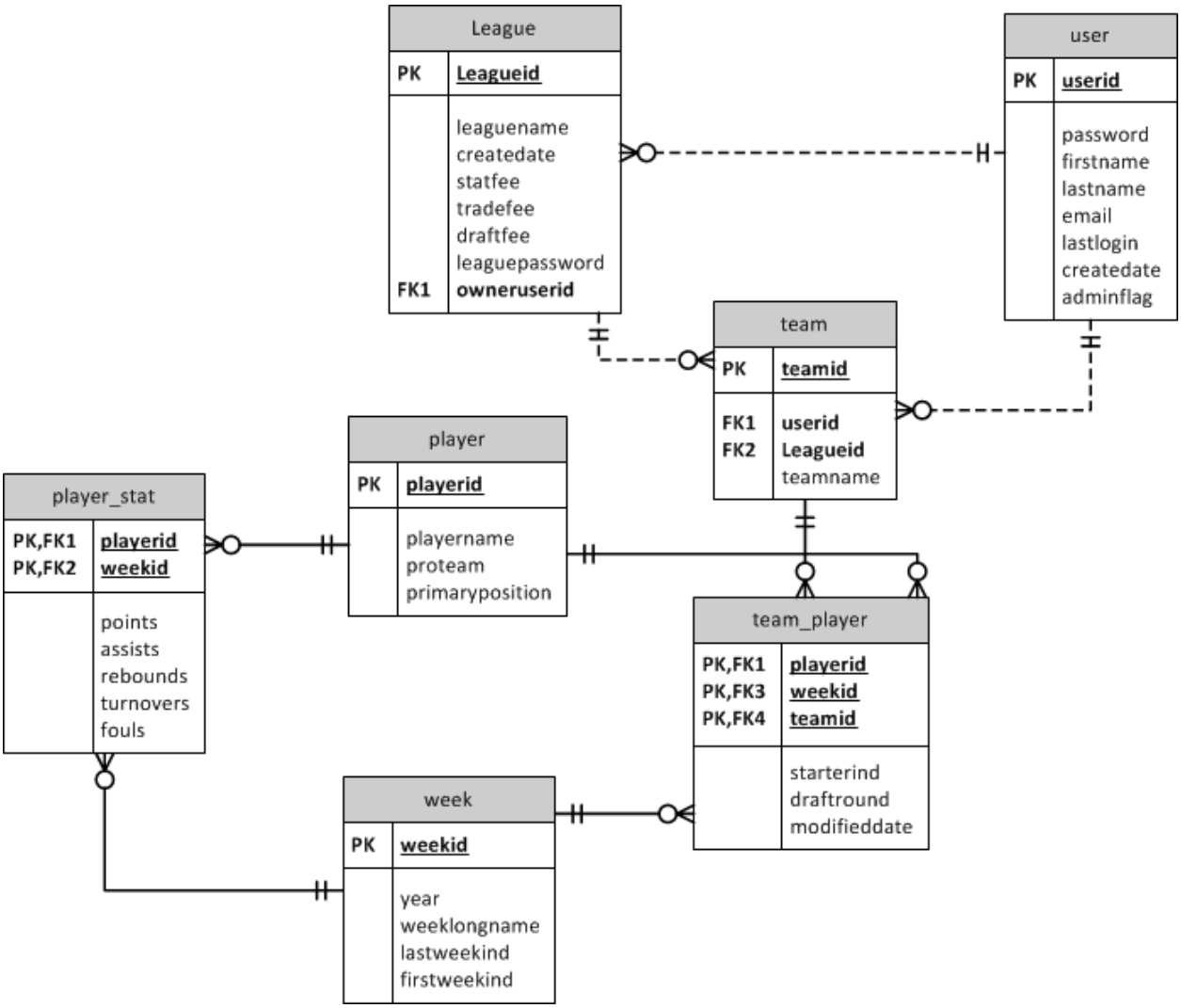
You should both describe your solution and provide a screen shot of the ER diagram. In addition, you should provide the Workbench file.

Description:

Because we're not dealing with any specific business problems right now, it's better to set the grain to be as detailed as possible. I defined my grain to be player's performance and team status per week per draft. The facts are the measurements we take according to our grain, which are startind, draftround, points, assists, rebounds, turnovers, fouls. The two dimension tables user and weeks are both role playing tables because the fact table references them twice by different foreign keys.


user
username VARCHAR(16)
email VARCHAR(255)
create_date DATETIME
password VARCHAR(32)
userid INT
firstname VARCHAR(45)
lastname VARCHAR(45)
lastlogin DATETIME
adminflag VARCHAR(45)
Indexes
team
teamid INT
teamname VARCH...
Indexes
Fact-Table
row_id INT
user_userid INT
league_owneruserid INT
team_teamid INT
Weeks_weekid INT
player_player_id INT
starterind TINYINT
draftround INT
modifieddate_id INT
points INT
assists INT
rebounds INT
turnovers INT
fouls INT
League_Leagueid INT
Indexes
Weeks
weekid INT
year YEAR
weeklongname VARCHA...
lastweekind TINYINT
firstweekind TINYINT
Indexes
player
player_id INT
playername VARCHAR(45)
proteam VARCHAR(45)
primaryposition VARCHAR(45)
Indexes
League
Leagueid INT
leaguename VARCHAR(45)
statfee DECIMAL
tradefee DECIMAL
draftfee DECIMAL
leaguepassword DOUBLE
Indexes

Appendix: Source Data Model



The entities in the database are described as follows:

Table	Definition
LEAGUE	Contains league-level information for each **Fantasy League**. The database can accommodate multiple leagues.
TEAM	Defines the **Fantasy Teams** that are in each league and their name.
TEAM_PLAYER	Defines what **Players** are on each **Fantasy Team** each week. A fantasy team is a list of players associated with one team in the league on any given week. Fantasy teams can change from week-to-week. starterind is an indicator starter player.
USER	Contains all the users (team owners) in the system.
WEEK	Contains all the valid weeks for playing fantasy soccer across all time. Lastweekind and firstweekind are indicators of whether this week is the first and last week of the season respectively. Weeklongname is the name of the week in long descriptive form (e.g. Week 3)
PLAYER	Contains a list of all the real-life soccer players that can be selected in the league. proteam records which team the player belongs to in the real world professional basketball.
PLAYER_STATS	Contains all the raw stats for each **Player** for a given week. Each non-key field is a numeric value for that week.