Homework #3 **Due: turned in by Mon 01/29/2020 before class**

Adil Ashish Kumar

(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 HW3 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).

Database is a collection of related data. Traditional databases are transactional. These are optimized to support well defined transaction requirements. It is hard to extract necessary information from databases in a way that is intuitive and fast for business users. In contrast data warehouses can be used for achieving the objective of decision optimized data storage so that information can be extracted to plan, make decisions and assess results.

Data warehouse is like a copy of transactional data structured specifically for query and analysis. Data warehouses are systems optimized for high performance queries so that users can answer questions that may require transactions to be queried and compressed into an answer set. Unlike a database, which is a transactional system, data warehouses do not deal with a single transaction at a time. Data warehouse contains same information as a database but package data differently to optimize query processing and improve understandability for business users. Data warehouses deal with analytical side of things and are based on OLAP approach. Data warehouses preserve historical context of data to monitor company performance over time.

Online Analytical processing (OLAP) is an approach to answer multidimensional analytical queries swiftly.

Online transaction processing (OLTP) facilitates and manages transactions, like data entry and retrieval. OLAP data content includes historical, derived and summarized data while OLTP content involves current data values. OLAP is optimized for complex queries while OLTP is optimized for transactions. OLAP data volume is in range of GB/TB,PB while OLTP data volume in MB/GB. OLAP access frequency is medium/low while OLTP access frequency is high. OLAP access type is read only while OLTP access type is read/update/delete. OLAP usage is random/ad hoc basis while OLTP usage is more predictable and repetitive. OLAP response time is seconds to minutes while OLTP response is in sub seconds. OLAP has a smaller number of users while OLTP has larger number of users.

Data mining is the process of analyzing large datasets in order to extract insights to drive business value. Data mining may require the use of databases or data warehouses to access the data. The trade off between using databases or datawarehouse for data mining depends on the format of data needed for data mining.

Business intelligence is the combination of analytics, data mining, data storage and infrastructure, and business practices to give organizations the framework to build insight and make data-driven decisions for the business. Data management and analysis practices are necessary to keep an organization's data clean, safe, and in a usable form for users across a company. BI technologies provide historical, current and predictive views of business operations. (wikipedia.com)

A data mart is a subset of a data warehouse oriented to a specific business line. Data marts contain repositories of summarized data collected for analysis on a specific section or unit within an organization, for example, the sales department(panoply.io)

ERP (Enterprise Resource planning) is integrated management of main business processes with help of sotrware and technology. It is typically a suite of integrated applications that is used by an organization to collect, manage, store and data. ERP provides an integrated and continuously updated view of core business processes using common databases maintained by a database management system (Wikipedia). ERP systems are like an application of data warehousing where data from transactional systems are all in one place to enable business users derive value from the data.

All other definitions were referred from Prof Panos Slides

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).

In normalized ER modeling, dimension tables are partially normalized while in dimensional modeling dimension tables are not normalized. Dimensional modeling have faster query/analysis performance while normalized ER modeling has performance degradation due to large no of joins required. ER modeling increases presentation complexity to users while dimensional modeling is more intuitive to users. Normalized ER modeling may reduce size of dimension table in comparison to dimensional modeling but it may not be a very useful feature.

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

Dimensional modeling results in more intuitive understanding for the end users while normalized ER modeling may not be very intuitive

Dimensional modeling provides better optimized query/analysis performance than normalized ER modeling

Dimensional modeling is more resilient to change in comparison to ER modeling as it easily accommodates change

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

1. Fact

Tells us about what a process is measuring. Most useful facts are numeric and additive as fact rows are generally accumulated to answer certain business questions

2. Grain

Grain tells us in business terms the level of detail associated with fact table measurments. It helps answer the question “how to describe a single row in the fact table”.

3. OLAP cube

Online analytical processing cube helps answer multi dimensional analytical queries swiftly. The data contains historical context and is in big sizes in order of GB/TB/PB

4. Snowflake schema

Snowflake schema normalizes the dimension tables in star schema. These structures are easier to maintain and reduce storage space required but are less intuitive to business users and have degraded query performance.

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

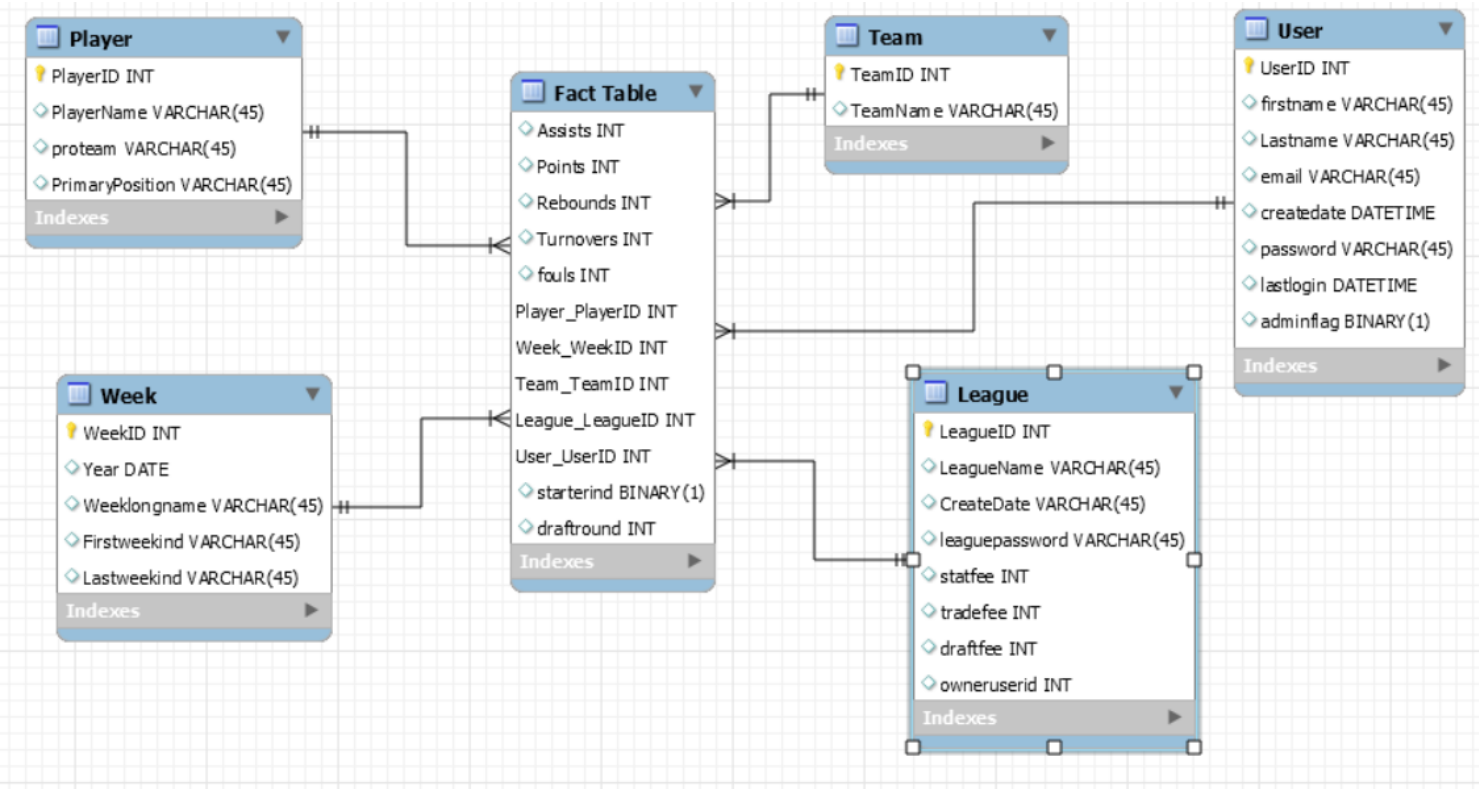
- **Who are the most drafted players across all leagues?**
- **Which user has the highest number of assists in the current season? (it means the user's players' assists while the user has them).**
- **Who are the most traded players in a particular fantasy league?**
- **How are teams ranked in a league 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:

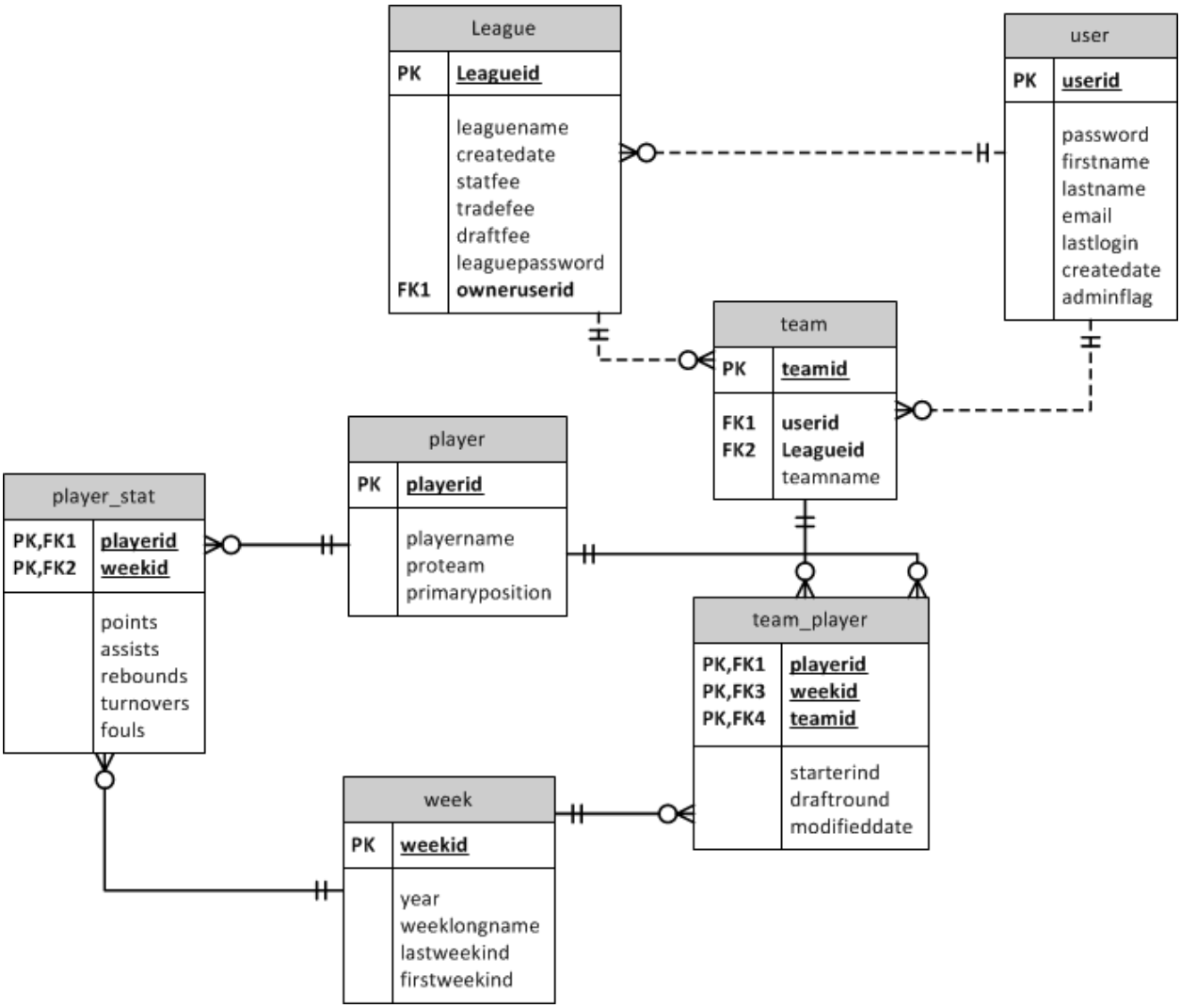
- **Step 1: What is the grain of the business process that we will model?**
- **Step 2: What are the facts?**
- **Step 3: What are the dimensions?**
- **Step 4: (Use MySQL Workbench) Draw an ER diagram with the fact and dimensions table. Identify the primary and foreign keys.**

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

Business process is a player playing each week across teams and leagues
Grain: a row would look like an individual player belonging to a user in a league and his performance in a game every week
Facts: points, assists, rebounds, fouls, turnovers, trades, drafts – all the facts that describe player performance for a week
Dimensions: league, players, team, week, user are the contraints/groupings associated with event


Player
PlayerID INT
PlayerName VARCHAR(45)
proteam VARCHAR(45)
PrimaryPosition VARCHAR(45)
Indexes
Fact Table
Assists INT
Points INT
Rebounds INT
Turnovers INT
fouls INT
Player_PlayerID INT
Week_WeekID INT
Team_TeamID INT
League_LeagueID INT
User_UserID INT
starterind BINARY(1)
draftround INT
Indexes
Team
TeamID INT
TeamName VARCHAR(45)
Indexes
User
UserID INT
firstname VARCHAR(45)
Lastname VARCHAR(45)
email VARCHAR(45)
createdate DATETIME
password VARCHAR(45)
lastlogin DATETIME
adminflag BINARY(1)
Indexes
Week
WeekID INT
Year DATE
Weeklongname VARCHAR(45)
Firstweekind VARCHAR(45)
Lastweekind VARCHAR(45)
Indexes
League
LeagueID INT
LeagueName VARCHAR(45)
CreateDate VARCHAR(45)
leaguepassword VARCHAR(45)
statfee INT
tradefee INT
draftfee INT
owneruserid INT
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.