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

Carl Xi

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

A. ERP

ERP stands for Enterprise Resource planning and refers to a system that integrate all day-to-day-business activities including but not limited to HR, sales, Finance and etc. It is typically used by corporations to monitor business activities in real time and make necessary adjustments. All the data used in ERPs are typically operational and transactional in nature, which is captured by complex operational systems, or OLTPs.

B. Database

- An organized electronic collection of stored data in forms such as tables for later access, modification, query or retrieval. Databases are typically designed to organize the data used for an organization's analytical needs in mind. A well-designed database should be easy to understand and query. Depending on the organization's needs, the database could take up various data structure forms, including OLTP or OLAP.

C. Data warehouse

- Data Warehousing is essentially a refined layer on top of databases that are designed and optimized for high-performance queries, queries that could take. They are large centralized data repositories that may contain data from several sources within a single organization. A good use of data warehousing is the storage of OLAP analytical results, where transactional data from disparate source systems are modified into historical or summarized analytical data.

D. Data mart

- Data Mart are designed within an organization's existing data warehousing system or independently and later converged with other independent data marts to form a data warehouse. An important purpose of data marts is to make the life of the people accessing the data easier, as it helps users narrow down on a single subject or a specific function within an organization.

E. OLAP

- OLAP stands for Online Analytical Processing and is a data structure form which is specifically designed with the goal of rapidly answering multi-dimensional analytical queries in mind. OLAPs have robust analytical capabilities and is geared towards analytical data, and is often used for the quick access, development, and/or management of dashboards, balance scorecards or analytical reports. OLAP systems provide read-only access to the data.

F. OLTP

- OLTP, which stands for Online Transactional Processing, is a data structure form which is specifically designed with the goal of facilitating and managing data entry and retrieval in mind. OLTP structure excels at managing high speed data transaction recording while keeping storage space utilization at a minimum. OLAPs are like the other side of the coin compared to OLTPs, as they are both data structures, but with different goals in mind. Unlike OLAPs, OLTP allow users to read, update and remove data entries.

G. Data Mining

- Data mining is an important step in data analytics and business intelligence and is the process of analytical prediction by finding anomalies, patterns or correlations within large datasets. Data mining is typically split into the 3 stages of initial exploration, model building/pattern identification/validation or verification, and finally deployment. Along with business intelligence which derives insights, data mining is typically used for organizational optimization and improvements such as increasing revenue, reducing costs, increasing customer satisfaction, managing/reducing risks and etc.

H. Business Intelligence

- Business intelligence is the combination of business analytics, data mining, data visualization, data tools, and data infrastructures. It's primary goal is to help organizations make more data-driven decisions by leveraging software services to transform data into actionable insights that guide strategic and tactical business decisions for organizations (Pratt & Fruhlinger, 2020), (tableau, 2020) (You can't come up with a better description than this). A typical case of business intelligence in practice is utilizing domain knowledge, data from data warehouses and further analytical work to derive actionable insights.

Bibliography

- What is business intelligence? Your guide to BI and why it matters. (n.d.). Tableau Software. Retrieved January 29, 2020, from https://www.tableau.com/learn/articles/business-intelligence
- What is data mining, predictive analytics, big data. (n.d.). Retrieved January 29, 2020, from http://www.statsoft.com/Textbook/Data-Mining-Techniques#mining
- What is erp? | oracle. (n.d.). Retrieved January 29, 2020, from https://www.oracle.com/applications/erp/what-is-erp.html
- What is data mining? (n.d.). Retrieved January 29, 2020, from https://www.sas.com/en_us/insights/analytics/data-mining.html
- What is data mart? - Bi glossary. (n.d.). Sisense. Retrieved January 29, 2020, from https://www.sisense.com/glossary/data-mart/
- Clinical data repository vs. Data warehouse: Which do you need? (2014, July 10). Health Catalyst. https://www.healthcatalyst.com/insights/clinical-data-repository-data-warehouse
- Fruhlinger, M. K. P. and J. (2019, October 16). What is business intelligence? Turning data into business insights. CIO. https://www.cio.com/article/2439504/business-intelligence-definition-and-solutions.html

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).

- While the ER Model can contain both logical and physical models, dimensional models can only include physical models. ER processes normalized data while dimensional models process de-normalized data. ER modeling is mostly used to remove data redundancy, ensure data consistency and express relationships between entities while dimensional modeling is typically used to capture critical measures and is viewed along dimensions. Overall, ER models are more useful for clerical users while dimensional models are more useful for business orientated users. As a final note, ER models are not mapped for creating schemas and uses current data while dimensional models are mapped for creating schemas and 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).

- Dimensional modeling is more flexible from the user perspective, and as several advantages over ER modeling for data warehouse design. First, the dimensional model is a predictable, standard framework. Report writers, query tools, and user interfaces can all make strong assumptions about the dimensional model to make the user interfaces more understandable and to make processing more efficient. (Kimballgroup, 1997) Additionally, the dimensional model also withstands unexpected changes in user behavior, as every dimension is equivalent. As such, all dimensions can be considered symmetrical entry points into the fact table. The logical design can thus be independent of expected query patterns. This aspect of symmetry is very important. (Kimballgroup, 1997) Lastly, the dimensional model is extensible in order to accommodate unexpected new data element and new design choices. It does so with minimal disturbance and revision needed.

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

1. Fact

- Fact is the performances measurements resulting from an organization's business process events. The measurement of a single event in the physical world represents one-on-one to a single row in the corresponding fact table. Examples of facts include sales quantity, per unit regular, discount, and net paid prices. It is typically stored in a fact table, where each row is represented by the grain, and a fact table combined with dimension tables becomes dimensional models. (Adamopoulos, 2020)

2. Grain

- The grain describes the level detail of the fact table measurements when conveyed in business terms. It answers the question "how do you describe a single row in the fact table?". We typically want the most finely grained data so that users can ask precise questions. (Adamopoulos, 2020)

3. OLAP cube

- OLAP stands for "OnLine Analytical Processing" and describes a swift approach to answering multi-dimensional analytical queries. (Adamopoulos, 2020). In data warehousing, data cubes describe multidimensional models that take advantage of inherent relationships in data to populate data in multidimensional matrices. (Elmasri & Shamkant, 2004) Together, OLAP cube describes using the swift computer-based analytical approach of OLAP to evaluate multi-dimensional data matrices in order to drive insights.

4. Snowflake schema

- The two common multinational schemas are the star schema and the snowflake schema. The snowflake schema is a variational of the star schema in which the dimensional tables from a star schema are organized into a hierarchy by normalizing them. In fact, some installations are normalizing data warehouses up to the third normal form so that they can access the data warehouse to the finest level of detail. (Elmasri & Shamkant, 2004)

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

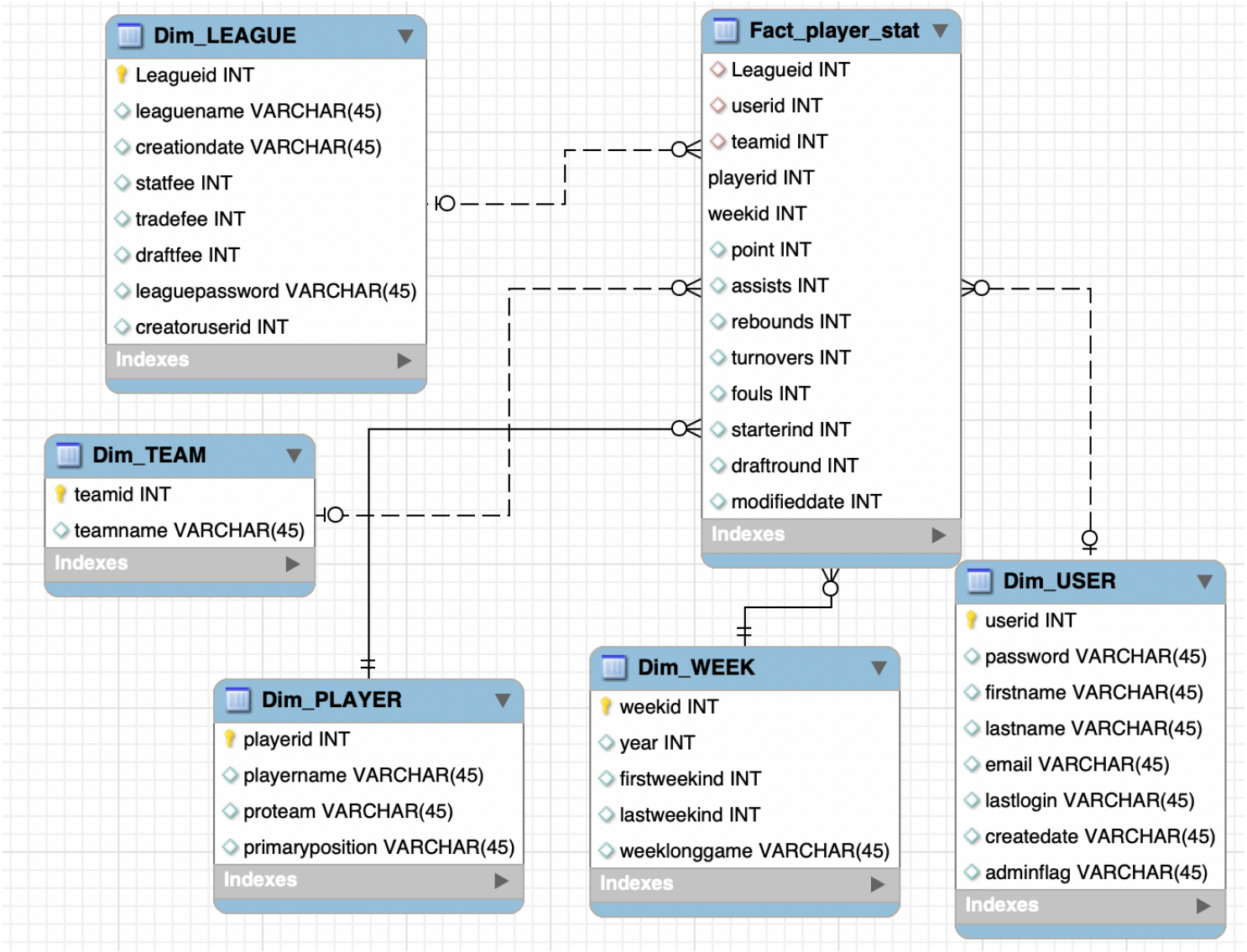
- **Who are the most drafted players across all leagues?**
- **Which user has the highest number of assists in the current season? (it means the user's players' assists while the user has them).**
- **Who are the most traded players in a particular fantasy league?**
- **How are teams ranked in a league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

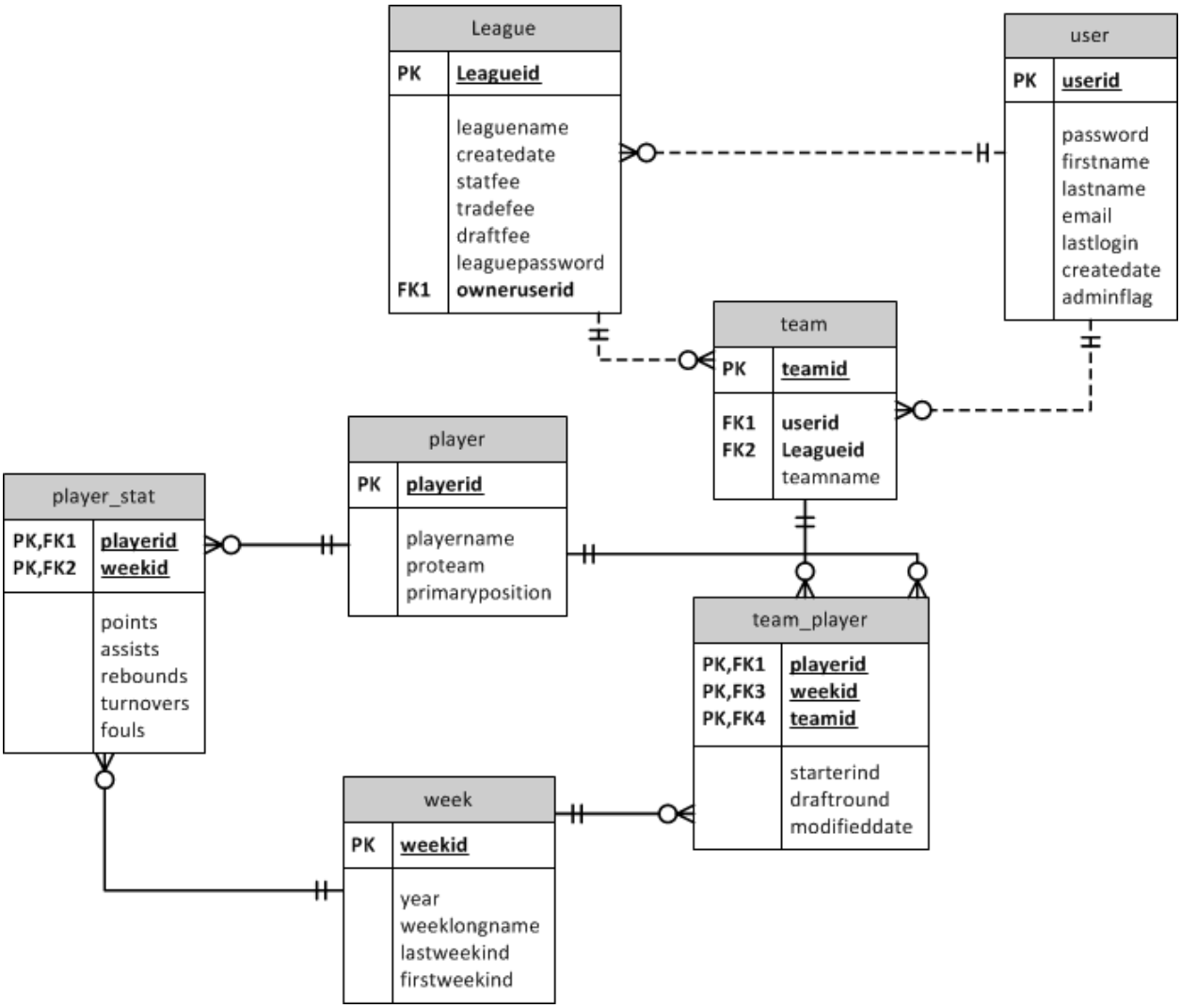
- **Step 1: What is the grain of the business process that we will model?**
- **Step 2: What are the facts?**
- **Step 3: What are the dimensions?**
- **Step 4: (Use MySQL Workbench) Draw an ER diagram with the fact and dimensions table. Identify the primary and foreign keys.**

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

Since we can't upload workbench files, you can find the workbench file here. Our ER diagram is very straightforward, with one fact and 5 dimensions tables. The grain of our business process is one player per team per league per week, as shown by the player's stats in the fact table (see the definition of grain and how it's related to the fact table above). The facts are simply player stats such as points, assists, rebounds, turnovers, fouls, starterind, draftround,and modified date. The dimensions are league the player is in, the team the player is in, the user the player is being used by, the week, and the player itself.


Dim_LEAGUE
Leagueid INT
leaguename VARCHAR(45)
creationdate VARCHAR(45)
statfee INT
tradefee INT
draftfee INT
leaguepassword VARCHAR(45)
creatoruserid INT
Indexes
Fact_player_stat
Leagueid INT
userid INT
teamid INT
playerid INT
weekid INT
point INT
assists INT
rebounds INT
turnovers INT
fouls INT
starterind INT
draftround INT
modifieddate INT
Indexes
Dim_TEAM
teamid INT
teamname VARCHAR(45)
Indexes
Dim_PLAYER
playerid INT
playername VARCHAR(45)
proteam VARCHAR(45)
primaryposition VARCHAR(45)
Indexes
Dim_WEEK
weekid INT
year INT
firstweekind INT
lastweekind INT
weeklonggame VARCHAR(45)
Indexes
Dim_USER
userid INT
password VARCHAR(45)
firstname VARCHAR(45)
lastname VARCHAR(45)
email VARCHAR(45)
lastlogin VARCHAR(45)
createdate VARCHAR(45)
adminflag VARCHAR(45)
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.