|  |  |
| --- | --- |
|  | **Managing Big Data** |

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

\_Xi (Athena) Li\_

(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).**

* ERP
  + ERP refers to Enterprise Resource Planning system. It provides an integrated and continuously updated view of core business processes using common databases maintained by a database management system (Wikipedia.org).
* Database
  + Database is an organized collection of data (Wikipedia.org).
* Data warehouse
  + A subject-oriented, integrated, nonvolatile, and time variant collection of data in support of management decisions (Toolkit, R. Kimball, 1996). A system used for reporting and data analysis, and is considered a core component of business intelligence (Wikipedia.org).
* Data mart
  + A data mart is the access layer of the data warehouse environment that is used to get data out to the users. The data mart is a subset of the data warehouse and is usually oriented to a specific business line or team. Whereas data warehouses have an enterprise-wide depth, the information in data marts pertains to a single department (Wikipedia.org).
* OLAP
  + OnLine Analytical Processing is an approach to answering multi-dimensional analytical (MDA) queries swiftly in computing. OLAP is part of the broader category of business intelligence. Databases configured for OLAP use a multidimensional data model, allowing for complex analytical and ad hoc queries with a rapid execution time. OLAP is typically contrasted to OLTP (online transaction processing), which is generally characterized by much less complex queries, in a larger volume, to process transactions rather than for the purpose of business intelligence or reporting. Whereas OLAP systems are mostly optimized for read, OLTP has to process all kinds of queries (read, insert, update and delete) (Wikipedia.org).
* OLTP
  + Online transaction processing (OLTP) is a class of information systems that facilitate and manage transaction-oriented applications, typically for data entry and retrieval transaction processing on a database management system (Wikipedia.org).
* Data Mining
  + Data mining is the computing process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. An essential process where intelligent methods are applied to extract data patterns. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. It bridges the gap from applied statistics and artificial intelligence (which usually provide the mathematical background) to database management by exploiting the way data is stored and indexed in databases to execute the actual learning and discovery algorithms more efficiently, allowing such methods to be applied to ever larger data sets (Wikipedia.org).
* Business Intelligence
  + Business Intelligence is a set of methodologies, processes, architectures, and technologies that transform raw data into meaningful and useful information used to enable more effective strategic, tactical, and operational insights and decision-making (Wikipedia.org).
  + Business Intelligence (BI) comprises the strategies and technologies used by enterprises for the data analysis of business information. BI technologies provide historical, current and predictive views of business operations(Wikipedia.org).
  + Often BI applications use data gathered from a data warehouse (DW) or from a data mart (Wikipedia.org).

**Bibliogrpgy**

**Enterprise resource planning. (2017, October 12). Retrieved October 15, 2017, from** [**https://en.wikipedia.org/wiki/Enterprise\_resource\_planning**](https://en.wikipedia.org/wiki/Enterprise_resource_planning)

**Database. (2017, October 11). Retrieved October 15, 2017, from https://en.wikipedia.org/wiki/Database**

**Data warehouse. (2017, October 13). Wikipedia, Wikimedia Foundation, Retrieved October 15, 2017, from en.wikipedia.org/wiki/Data\_warehouse.**

**Data mart. (2017, October 10). Retrieved October 15, 2017, from** [**https://en.wikipedia.org/wiki/Data\_mart**](https://en.wikipedia.org/wiki/Data_mart)

**Online analytical processing. (2017, October 10). Retrieved October 15, 2017, from**

**https://en.wikipedia.org/wiki/Online\_analytical\_processing**

**Online transaction processing. (2017, October 12). Retrieved October 15, 2017, from**

[**https://en.wikipedia.org/wiki/Online\_transaction\_processing**](https://en.wikipedia.org/wiki/Online_transaction_processing)

**Data mining. (2017, October 12). Retrieved October 15, 2017, from**

**https://en.wikipedia.org/wiki/Data\_mining**

**Business intelligence. (2017, October 12). Retrieved October 15, 2017, from**

**https://en.wikipedia.org/wiki/Business\_intelligence**

**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. Dimension models are quite different from 3NF models which seek to remove data redundancies.
2. A dimensional model contains the same information as a normalized model, but packages the data in a format that delivers user understandability, query performance, and resilience to change.
3. In a star schema, each business process is represented by a dimensional model that consists of a fact table containing the event’s numeric measurements. The fact table is surrounded by a halo of dimension tables that contain the textual context that was true at the moment the event occurred.
4. The ER Modeling is a transactional modeling process, while dimensional modeling is a analytical modeling process.
5. Star schema is de-normalized dimension tables

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

1. Transactional models, while efficient for transaction processing, are not good for analytics. So we use dimensional modeling to organize the data so that it will be pulled out efficiently.
2. Dimensional modeling is preferred for presenting analytic data because it delivers data that’s understandable to the business users and it delivers fast query performance.
3. The complexity of users’ unpredictable queries using normalized ER modeling overwhelms the database optimizers

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

**1. Fact**

The fact table in a dimensional model stores the performance measurements resulting from an organization’s business process events

**2. Grain**

Grain conveys in business terms the level of detail associated with the fact table measurements (an individual fact table row).

**3. OLAP cube**

OLAP cube refers to multi-dimensional array of data. OLAP (OnLine Analytical Processing) is an approach to answer multi-dimensional analytical queries swiftly

**4. Snowflake schema**

Snowflaking is a method of normalizing the dimension tables in a STAR schema.

**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**

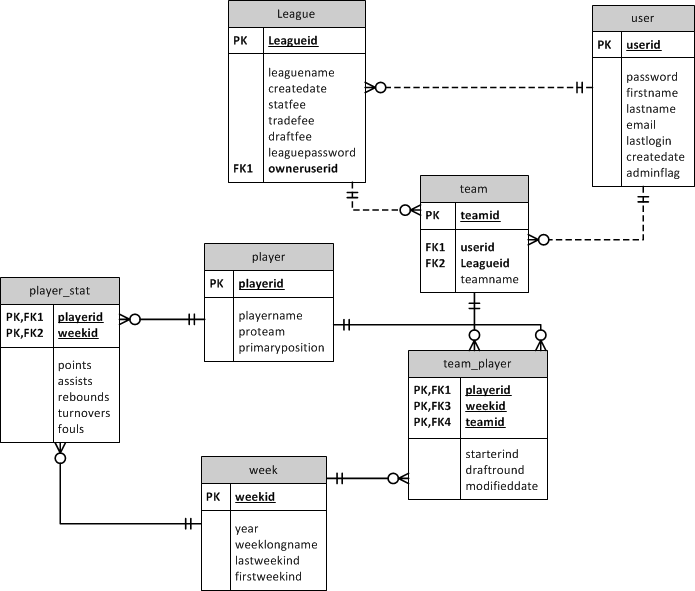
* Who are the most drafted players across all leagues?
* Which user has the highest number of assists in the current season? (it means the user’s players’ assists while the user has them).
* Who are the most traded players in a particular fantasy league?
* How are teams ranked in a league in terms of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

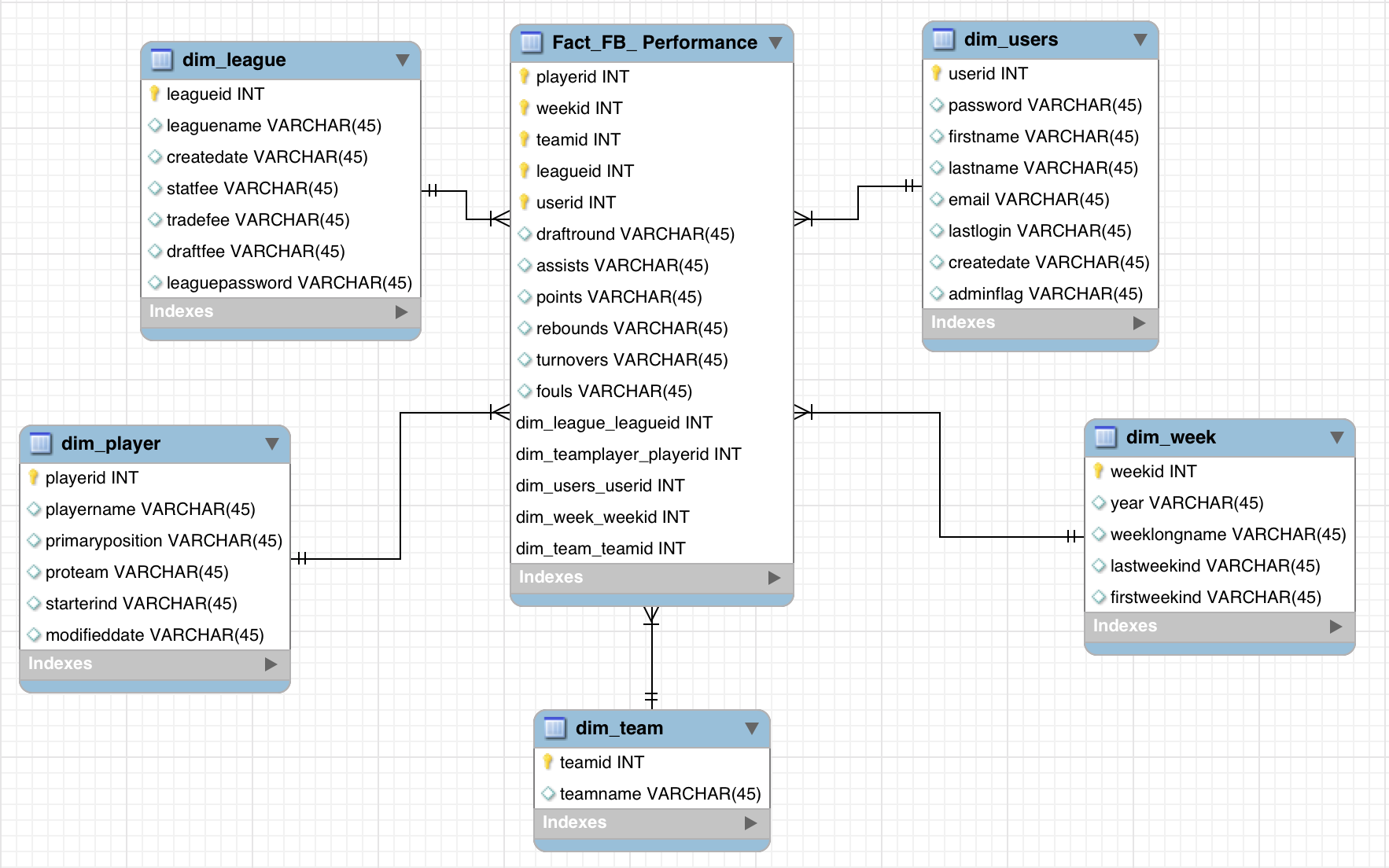
**Appendix: Source Data Model**



**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. starterindis 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. |

* **What is the grain?**
  + Based on the question, we need the most detailed information for the data warehouse, which means we’ll have 1 Fantasy Basketball player’s performance on each week in a given team per row. So, the grain would be playerid-weekid-teamid.
* **What are the facts?** 
  + **Draftround**
  + **Assists**
  + **Points**
  + **Rebounds**
  + **Turnovers**
  + **Fouls**
* **What are the dimensions?** 
  + **League**
  + **Player**
  + **Team**
  + **Users**
  + **Week**

****