Assignment 2
CIS 453/553 Data Mining, Spring 2017
due 11:59 pm, Friday April 28th

1. A data warehouse for previous NCAA football games has five dimensions (date, game, location, player, and spectator), and two measures (count and charge). Charge is what a spectator pays when watching a game in a given date and location. The spectators can be students, faculty, adults, seniors (older than 65), children (younger than 12), with each category having its own charge rate. The player information includes name, age, height, weight, position, the numbers of Receiving Yards, Rushing Yards, Interceptions, Fumbles, Tackles, Touch Downs, and Field Goals in a game.

(a) Draw a star schema for the data warehouse.

(b) If we treat spectators and players as persons who all have name, address, age. For example, players in one game can be spectators of another game. Please design a new schema to represent this.

(c) Starting with the base cuboid [date, game, location, player, spectator], what specific OLAP operations (e.g., roll-up from quarter to year) should one perform in order to list the total charge paid by all students at Autzen Stadium in the fall of 2016 and Justin Herbert was one of players?

(d) Using a starnet query model to represent your design in (a) and the query in (c).

2. We talked about virtual data warehouse and data mediator in the class. Are they exactly the same thing? If not, what's the difference? Whether the virtual data warehouse can support OLAP mining (OLAM)?

3. Suppose that a data warehouse contains 20 dimensions, each with about five levels of granularity. Users are mainly interested in four particular dimensions, each having three frequently accessed levels for rolling up and drilling down. How would you design a data cube structure to support this preference efficiently?

4. Suppose a bank database includes following attributes to describe customers: name, age, gender, address, phone#, credit-ranking (good and bad), year-income, job title (student, engineer, professor etc..). Based on your selected schema in Problem 4 of assignment 1 (Generalized relation), write a DMQL query to compare the general properties between customers who have good or bad credit-ranking. (Note, you do not need to exactly follow the DMQL syntax if you do not know SQL.)

To turn in by paper version: Ask Cheri to put your answers to Prof. Dejing Dou's mailbox.

To turn in by emails: We prefer that you send in a pdf file. If you are using Word, you should be able to convert your word file to a pdf file.