Data Warehousing Assignment-2

This problem set consists of two data modeling scenarios. You will be asked to analyze the strengths and weaknesses of some design alternatives for each scenario. Short answers are fine – one or two paragraphs per question would be an appropriate length.

Scenario I

In this scenario, we are interested in modeling student enrollment in Stanford courses. We would like to answer questions such as:

- Which courses are most popular? Which instructors are most popular?
- Which courses are most popular among graduate students? Undergraduates?
- Are there courses for which the assigned classrooms is too large or too small?

We are planning to have a course enrollment fact table with the grain of one row per student per course enrollment. In other words, if a student enrolls in 5 courses there will be 5 rows for that student in the fact table. We will use the following dimensions: Course, Department, Student, Term, Classroom, and Instructor. There will be a single fact measurement column, Enrollment Count. Its value will always be equal to 1.

We are considering several options for dealing with the Instructor dimension. Interesting attributes of instructors include First Name, Last Name, Title (e.g. Assistant Professor), Department, and Tenured Flag. The difficulty is that a few courses (less than 5%) have multiple instructors. Thus it appears we cannot include the Instructor dimension in the fact table because it doesn't match the intended grain. Here are the options under consideration:

OptionA

ModifytheInstructordimensionbyaddingspecialrowsrepresentinginstructorteams.Forexample, CS276ais taught by Manning and Raghavan, so there will be an Instructor row representing "Manning/Raghavan" (as well as separate rows for Manning and Raghavan, assuming that they sometimes teach courses as sole instructors). In this way, the Instructor dimension becomes true to the grain and we can include it in the fact table.

Option B

Change the grain of the fact table to be one row per student enrollment per course per instructor. For example, there will be two fact rows for each student enrolled in CS 276a, one that points to Manning as an instructor and one that points to Raghavan. However, each of the two rows will have a value of 0.5 in the EnrollmentCount field instead of a value of 1, in order to allow the fact to aggregate properly. (Enrollments are "allocated" equally among the multiple instructors.)

Option C

Create two fact tables. The first has the grain of one row per student enrollment per course and doesn't include the Instructor dimension. The second has the grain of one row per student enrollment per course per instructor and includes the Instructor dimension (as well as all the other dimensions). Unlike Option B, the value of Enrollment Count will be 1 for all rows in the second fact. Tell warehouse users to use the second fact table for queries involving attributes of the instructor dimension and the first fact table for all other queries.

Please answer the following questions.

Question 1. What are the strengths and weaknesses of each option?

Answer.

strengths:

Option A. This option allows the Instructor dimension to be included in the fact table, which will enable users to easily query the data for information about specific instructors. It also allows the grain of the fact table to remain at one row per student per course, which will make it easy to aggregate data for analysis.

Option B. This option allows for the inclusion of the Instructor dimension in the fact table and enables accurate representation of enrollments for each instructor. It also allows for the possibility of querying data at the student-course-instructor grain, which may be useful for some analysis.
Option C. This option allows for the creation of two separate fact tables, one for general enrollment data and one for data at the student-course-instructor grain. This allows users to choose the appropriate fact table for their specific analysis needs and ensures that the data is accurately represented in each table.
Weaknesses:
Option A. This option requires the creation of special rows in the Instructor dimension to represent instructor teams, which could make the dimension more complex and potentially harder to understand and use. Additionally, this option may not accurately represent the actual enrollment of each instructor, as the enrollment will be split equally among the instructors in a team.
Option B. This option requires changing the grain of the fact table to one row per student per course per instructor, which may make it more difficult to aggregate data for analysis at higher levels. It also requires the use of fractional values in the Enrollment Count field, which may be confusing for users.
Option C. This option requires the creation and maintenance of two separate fact tables, which may increase complexity and the burden on the warehouse. It may also require users to remember to use the appropriate fact table for their queries, which could lead to errors or incorrect results if they use the wrong table.

Question 2. Which option would you choose and why?
Answer. Given the options presented, I would choose Option C. This option allows for the creation of two separate fact tables, one with the Instructor dimension and one without, which enables users to choose the appropriate table for their analysis needs. This ensures that the data is accurately represented and avoids the need for special rows or fractional values in the fact table.

Question 3. Would your answer to Question 2 be different if the majority of classes had multiple instructors? How about if only one or two classes had multiple instructors? (Explain your answer.)
Answer. If the majority of classes had multiple instructors, I would still choose Option C. This option would allow for the creation of two separate fact tables that accurately represent the data for classes with single and multiple instructors. If only one or two classes had multiple instructors, I may consider Option A or B as well. However, I would still ultimately choose Option C as it allows for the greatest flexibility in querying and analysis, while still ensuring that the data is accurately represented.

Question 4. [OPTIONAL] Can you think of another reasonable alternative design besides Options A, B, and C? If so, what are the advantages and disadvantages of your alternative design?
Answer. An alternative design that could be considered is to include the Instructor dimension in the fact table, but to also include a separate column for instructor allocation, which would represent the fraction of the enrollment that each instructor is responsible for. This would allow for the inclusion of the Instructor dimension in the fact table while still accurately representing enrollments for each instructor. However, this design may also increase the complexity of the fact table and could potentially be confusing for users.

Scenario II

In this scenario, we are building a data warehouse for an online brokerage company. The company makes money by charging commissions when customers buy and sell stocks. We are planning to have a Trades fact table with the grain of one row per stock trade. We will use the following dimensions: Date, Customer, Account, Security (i.e. which stock was traded), and TradeType.
The company's data analysts have told us that they have developed two customer scoring techniques that are used extensively in their analyses.

- Each customer is placed into one of nine Customer Activity Segments based on their frequency of transactions, average transaction size, and recency of transactions.
- Each customer is assigned a Customer Profitability Score based on the profits earned as a result of that customer's trades. The score can be either 1,2,3,4, or 5, with 5 being the most profitable.

These two scores are frequently used as filters or grouping attributes in queries. For example:

- How many trades were placed in July by customers in each customer activity segment?
- What was the total commission earned in each quarter of 2003 on trades of IBM stock by customers with a profitability score of 4 or 5?

There are a total of 100,000 customers, and scores are recalculated every three months. The activity level or profitability level of some customers changes over time, and users are very interested in understanding how and why this occurs.

We are considering several options for dealing with the customer scores:

Option A

The scores are attributes of the Customer dimension. When scores change, the old score is overwritten with the new score (Type 1 Slowly Changing Dimension).

Option B

The scores are attributes of the Customer dimension. When scores change, new Customer dimension rows are created using the updated scores (Type 2 Slowly Changing Dimension).

Option C

The scores are stored in a separate CustomerScores dimension which contains 45 rows, one for each combination of activity and profitability scores. The Trades fact table includes a foreign key to the CustomerScores dimension.

Option D

The scores are stored in a CustomerScores outrigger table which contains 45 rows. The Customer dimension includes a foreign key to the outrigger table (but the fact table does not). When scores change, the foreign key column in the Customer table is updated to point to the correct outrigger row.

Please answer the following questions.

Question 5. What are the strengths and weaknesses of each option?

Answer.

Option A

Strengths: This option allows for the scores to be stored as attributes of the Customer dimension, which will make it easy for users to filter or group by scores in their queries. It also avoids the need for additional dimensions or outrigger tables, which may simplify the overall design of the warehouse.

Weaknesses: This option involves overwriting old scores with new scores when they change, which means that the warehouse will not retain a history of changes to customer scores. This may make it difficult for users to understand how and why a customer's score changes over time.

Option B

Strengths: This option allows for the retention of a history of changes to customer scores, which may be useful for analysis and understanding how scores change over time. It also allows for the scores to be stored as attributes of the Customer dimension, which will make it easy for users to filter or group by scores in their queries.

Weaknesses: This option requires the creation of new Customer dimension rows each time a score changes, which may increase the complexity and size of the dimension. It may also make it more difficult for users to identify the current score for a customer, as they will have to determine which row represents the most recent score.

Option C

Strengths: This option allows for the scores to be stored in a separate CustomerScores dimension, which may make it easier for users to filter or group by scores in their queries. It also allows for the retention of a history of changes to customer scores, as the fact table includes a foreign key to the CustomerScores dimension rather than the scores themselves.

Weaknesses: This option requires the creation of a separate CustomerScores dimension with 45 rows, which may increase the complexity of the warehouse design. It also requires the use of foreign keys in the fact table, which may make it more difficult to perform certain types of analysis.

Option D

Strengths: This option allows for the scores to be stored in an outrigger table, which may make it easier for users to filter or group by scores in their queries. It also allows for the retention of a history of changes to customer scores, as the Customer dimension includes a foreign key to the outrigger table rather than the scores themselves.
Weaknesses: This option requires the use of an outrigger table, which may increase the complexity of the warehouse design. It also requires the use of foreign keys in the Customer dimension, which may make it more difficult to perform certain types of analysis.

Question 6. Which option would you choose and why?
Answer.
Given the options presented, I would choose Option B. This option allows for the retention of a history of changes to customer scores, which may be useful for analysis and understanding how scores change over time. It also allows for the scores to be stored as attributes of the Customer dimension, which will make it easy for users to filter or group by scores in their queries.

Question 7. Would your answer to Question 6 be different if the number of customers and/or the time interval between score recalculations was much larger or much smaller? (Explain your answer.)
Answer.
If the number of customers or the time interval between score recalculations was much larger, I would still choose Option B. This option allows for the retention of a history of changes to customer scores, which may be more useful for analysis and understanding how scores change over time when dealing with a larger number of customers or more frequent score recalculations.
If the number of customers or the time interval between score recalculations was much smaller, I may consider Option A as well. In this case, the increased complexity and size of the Customer dimension associated with Option B may not be justified, as the number of changes to customer scores is likely to be much smaller.