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, EnrollmentCount. Its value will always be equal to 1.

We are considering several options for dealing with the Instructor dimension. Interesting attributes of instructors include FirstName, LastName, Title (e.g. Assistant Professor), Department, and TenuredFlag. 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

Option B

Option C

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.

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.)

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

EnrollmentCount 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?

Question 2. Which option would you choose and why?

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.)

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?

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.
- EachcustomerisassignedaCustomerProfitabilityScorebasedontheprofitsearnedasaresultofthatcustomer'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:

OptionA Option B Option C

Option D

ThescoresareattributesoftheCustomerdimension.Whenscoreschange,theoldscoreisoverwrittenwiththe new score (Type 1 Slowly Changing Dimension).

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).

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

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?

Question 6. Which option would you choose and why?

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.)

Question 8. [OPTIONAL] Can you think of another reasonable alternative design besides Options A, B, C, and D? If so, what are the advantages and disadvantages of your alternative design?

Scenario-1

Question 1: Strengths and Weaknesses of Each Option

Option A (Modify Instructor Dimension):

Strengths: Maintains the original grain, allowing inclusion of the Instructor dimension. Handles instructor teams but still provides individual details.

Weaknesses: Increased complexity in queries involving individual instructors. Requires additional effort in maintaining and updating instructor data.

Option B (Change Fact Table Grain):

Strengths: Detailed tracking of each instructor's involvement. Straightforward for querying individual instructors.

Weaknesses: Introduces fractional values in EnrollmentCount, potentially complicating aggregation. Less intuitive for other analysis purposes.

Option C (Two Fact Tables):

Strengths: Separates concerns, providing flexibility and simplicity. Allows users to choose the appropriate table based on the type of query.

Weaknesses: Requires users to switch between tables for comprehensive analysis. May impact performance due to multiple tables.

Question 2: Which Option and Why?

I would choose Option C because it strikes a balance between maintaining the original grain for general enrollment information and providing a dedicated fact table for instructor-specific details. This separation simplifies queries and allows users to choose the appropriate table based on their specific analytical needs.

Question 3: Differences with Multiple Instructors?

If the majority of classes had multiple instructors, Option C would likely remain preferable as it handles such scenarios by providing a dedicated table for instructor details. Option B may become less practical due to increased fractional values and potential confusion.

If only one or two classes had multiple instructors, the impact on Option B might be minimal, but the choice would still depend on specific priorities and use cases. Option A might also be considered if detailed information about instructor teams is crucial.

Question 4: Optional Alternative Design

An alternative design could involve creating a separate fact table specifically for instructor information, linked to the main enrollment fact table. This maintains simplicity in the main fact table while allowing users to join with the instructor-specific fact table for detailed analysis. The advantage is simplicity in the main fact table, but the disadvantage is the need for additional joins for comprehensive analysis involving instructors.

Scenario-2

Question 5: Strengths and Weaknesses of Each Option

Option A (Type 1 Slowly Changing Dimension):

Strengths: Simplicity in data model; easy to implement. Requires less storage.

Weaknesses: Loss of historical data; unable to track changes in customer scores over time.

Option B (Type 2 Slowly Changing Dimension):

Strengths: Maintains historical changes in customer scores; allows for tracking changes over time.

Weaknesses: Increased storage requirements due to multiple rows for the same customer.

Option C (Separate CustomerScores Dimension):

Strengths: Preserves historical changes; provides flexibility for different combinations of scores. Avoids impact on the main Customer dimension.

Weaknesses: Adds complexity with an additional dimension; may require more complex queries.

Option D (CustomerScores Outrigger Table):

Strengths: Separates scores into an outrigger table, avoiding direct impact on the main Customer dimension. Allows for historical tracking.

Weaknesses: Adds complexity with an outrigger table; requires additional maintenance when scores change.

Question 6: Which Option and Why?

I would choose Option C because it provides a balance between preserving historical changes, flexibility, and avoiding direct impact on the main Customer dimension. This option allows for efficient querying and analysis of customer scores without complicating the primary customer data.

Question 7: Differences with Larger/Smaller Customer Base or Time Intervals?

Larger Customer Base: Option C may remain a good choice as it efficiently handles a larger number of customers by storing scores in a separate dimension. Options A and B may become less practical due to potential performance issues and increased storage requirements.

Smaller Customer Base: The impact on options might be minimal, but Options A and B could be more acceptable due to lower storage and simplicity. However, Option C still provides flexibility and maintainability.

Time Intervals: If the time interval between score recalculations is much larger, Options A and B might be more acceptable due to reduced frequency of updates. However, Option C's flexibility still makes it a viable choice.

Question 8: Optional Alternative Design

An alternative design could involve a hybrid approach, where the CustomerScores dimension is maintained separately, but only the significant changes are tracked in the main Customer dimension (similar to a Type 1 dimension). This balances the need for historical tracking and simplicity in the primary dimension. However, it would require careful consideration of what changes are significant enough to be included in the main dimension.