Q)

Data Warehousing Assignment

This problem set consists of two data modeling scenarios. You will be asked to analyze the strengths and weak nesses 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?

A)

Scenario I - Student Enrollment in Stanford Courses:

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

Option A:

Strengths:

- The Instructor dimension can be included in the fact table, maintaining the intended grain.

- It allows for capturing multiple instructors for courses that have them.

Weaknesses:

- It may lead to increased complexity in querying and reporting, as the instructor names need to be parsed and processed.

- It requires additional rows for instructor teams, potentially increasing the size of the dimension table.

Option B:

Strengths:

- The fact table maintains the intended grain of one row per student enrollment per course.

- It allows for capturing multiple instructors and properly aggregating enrollments.

Weaknesses:

- It introduces fractional values in the EnrollmentCount field, which might lead to complications in calculations and comparisons.

- The fact table becomes larger due to duplicating rows for each instructor associated with a course.

Option C:

Strengths:

- It provides two separate fact tables, each optimized for different types of queries.

- The first fact table without the Instructor dimension is suitable for general queries.

- The second fact table with the Instructor dimension is useful for queries involving instructor attributes.

Weaknesses:

- It increases the complexity of the data model and the maintenance of multiple fact tables.

- It may require additional efforts to educate and guide warehouse users on which fact table to use for specific queries.

Question 2. Which option would you choose and why?

The choice of option depends on the specific requirements and priorities of the analysis. If the analysis heavily relies on instructor-related attributes and metrics, Option C would be a suitable choice as it provides dedicated fact tables for different types of queries. However, if the instructor-related attributes are not critical for most of the analysis and the main focus is on course popularity, enrollment patterns, and classroom suitability, Option A could be a simpler and more straightforward approach.

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.)

If the majority of classes had multiple instructors, Option B would become more favorable as it handles multiple instructors more efficiently. It ensures that enrollments are properly allocated among the instructors, allowing for accurate analysis of individual instructor performance and the impact of multiple instructors on course popularity.

If only one or two classes had multiple instructors, the choice between Option A and Option B would depend on the importance of capturing individual instructor contributions. If it is essential to distinguish between individual instructors, even for those few classes, Option B would still be preferred. However, if capturing the specific instructor for those classes is not critical, Option A could be simpler to implement.

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?

One alternative design could be to create a bridge table between the fact table and the Instructor dimension. This bridge table would establish a many-to-many relationship between enrollments and instructors. Each row in the bridge table would represent an enrollment linked to one or more instructors. This design allows for capturing multiple instructors for each enrollment without duplicating fact table rows or modifying the grain of the fact table. However, it introduces additional complexity in querying and joins, as the bridge table needs to be used to retrieve instructor-related information.