Datawarehouse assignment – 02:

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:

Option A

Modify the Instructor dimension by adding special rows representing instructor teams. For example, CS276a is 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 enrolment 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 Enrolment Count field instead of a value of 1, to allow the fact to aggregate properly. (Enrolments are “allocated” equally among the multiple instructors.)

Option C

Create two fact tables. The first has the grain of one row per student enrolment per course and doesn’t include the Instructor dimension. The second has the grain of one row per student enrolment per course per instructor and includes the instructor dimension (as well as all the other dimensions). Unlike Option B, the value of Enrolment 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?

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?

Option A:

Strengths:

* It allows the instructor dimension to be included in the fact table, which will enable users to query the data for information quickly about specific instructors.
* It also allows the grain of the fact table to remain at one row per student per course, in which aggregate data for analysis will be made easy.

Weaknesses:

* 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.
* It 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:

Strengths:

* 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.

Weaknesses:

* 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:

Strengths:

* 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:

* This option requires the creation and maintenance of two separate facttables, 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?

* I would prefer option C. As we will create two facts tables, one for instructor dimensions and

One without instructor. This enables users to represent and query data according to their needs.

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 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 the flexibility in querying and analysis, while still ensuring that the data is accurately represented.

Question 2:

Option A: The scores are attributes of the Customer dimension.

Strengths: Simple implementation, as it requires only updating the existing Customer dimension row. It maintains a history of changes but only for the current row.

Weaknesses: Loss of historical data as the old score is overwritten. It does not provide a complete view of the changes over time, limiting analysis on score trends.

Option B: The scores are attributes of the Customer dimension (Type 2 Slowly Changing Dimension).

Strengths: Maintains a complete history of score changes over time by creating new rows in the Customer dimension. Allows analysis of score trends and the ability to understand how and why changes occur.

Weaknesses: Increased storage requirements due to the creation of new rows for each score change. Querying against the Customer dimension may involve additional complexity when handling multiple rows for the same customer.

Option C: The scores are stored in a separate Customer Scores dimension.

Strengths: Provides a dedicated dimension to store scores, allowing efficient storage and management. Enables easy filtering and grouping by scores in queries.

Weaknesses: Requires maintaining a separate dimension table, which adds complexity to data management. May result in redundant data if some score combinations are not used by any customer.

Option D: The scores are stored in a Customer Scores outrigger table.

Strengths: Keeps the scores separate from the Customer dimension, reducing data duplication. Allows for efficient storage and easy association between customers and scores.

Weaknesses: The fact table does not directly include the foreign key to the outrigger table, which can complicate certain queries. Updating the foreign key in the Customer dimension requires extra steps to ensure data consistency.

Question 6.

Option B provides the best solution for tracking and analyzing changes in customer scores over time. It maintains whole history of score changes by creating new rows in the Customer dimension. This enables other teams to understand the trends and reasons behind changes in customer activity and profitability levels. It also provides flexibility in querying and reporting on different score combinations.

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.)

If the number of customers is much larger, Option C (separate Customer Scores dimension) may become more favorable. With a significantly larger customer base, maintaining a separate dimension table for scores can be more efficient in terms of storage and management. It allows for easy filtering and grouping by scores without creating redundant data for unused score combinations.

On the other hand, if the time interval between score recalculations is much smaller, Option A (Type 1 Slowly Changing Dimension) may be more suitable. With frequent score updates, storing historical changes for every score recalculation using Option B (Type 2 Slowly Changing Dimension) could result in excessive data growth and increased complexity in managing the dimension table.

=====================================================================================