Data Warehousing Assignment

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 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:

Option A

Option B

Option C

Modify the Instructor dimension by adding special rows representing instructor teams. For example,CS276ais taught by Manning and Raghavan, so there will be an Instructor row representing "Manning/Raghavan" (as well as separaterows 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 Enrollment Count 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

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?

Option	**Strengths**	**Weaknesses**
Option A	Suitable for to run queries related to specifically Student enrollment data	Not suitable for analyzing instructor specific queries
Option B	Best suitable for querying students' enrolment data and instructors' data	Requires additional transformation and aggregation
Option C	Minimal space and easy to maintain	May raise conflicts for users and more data shuffling required

Question 2. Which option would you choose and why?

Option _ B is recommenced as it serves the business requirements well with minimal storage

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.)

No, as the Option B serves the for Question 3 scenario also.